Postcode Anonymity Analysis

Adam Hardy

May 2021

1 Introduction

In the US, 5-digit Zip codes are usually rounded to 3-digits when anonymising health data, so knowledge of the Zip code doesn't allow small groups to be identified. Even then, there are some 3-digit codes that have fewer than 20,000 residents, and the advice is to lump these together under a new code (000). Looking forward to how GDPR may affect data handling in the UK, is it possible to use a similar approach here?

UK postcodes can be in one of six formats and broken down into seven components that describe geographic areas (Table 1). Given how UK postcodes are constructed, we can't simply truncate postcodes to three characters as is done for USA zip codes as this will result in postcode areas being combined together in ways which do not have real geographic meaning. For example, 'NE3 1ED' and 'NE35 2FG' are both valid postcodes. Truncating both of these to the first three characters would put these both in the 'NE3' group. They should instead both be in an 'NE' area group or separate 'NE3' and 'NE35' district groups (or another chosen postcode component group).

Below, we take a UK postcode and population dataset[1] and perform an anonymity analysis of the postcode geographic components on the four population groups present in the dataset: total population, male population, female population and occupied households.

Postcode Format	Outward Code	Inward Code	Area	District	Sub-District	Sector	Unit
AA9A 9AA	AA9A	9AA	AA	AA9	AA9A	AA9A 9	AA
A9A 9AA	A9A	9AA	A	A9	A9A	A9A 9	AA
A9 9AA	A9	9AA	A	A9	**N/A**	A9 9	AA
A99 9AA	A99	9AA	A	A99	**N/A**	A9 9	AA
AA9 9AA	AA9	9AA	AA	AA9	**N/A**	AA9 9	AA
AA99 9AA	AA99	9AA	AA	AA99	**N/A**	AA99 9	AA

Table 1: UK postcode formats and geographic components. 'A' represents an alphabetic character, '9' represent a numeric character. Obtained from `https://ideal-postcodes.co.uk/guides/uk-postcode-format`.

[1] `http://www.nomisweb.co.uk/output/census/2011/Postcode_Estimates_Table_1.csv`

2 Anonymity Analysis

2.1 Anonymity Threshold

For our anonymity analysis we are most concerned about small groups, as they are most likely to contain non-anonymous data. We will define an anonymity threshold, which is the smallest group size that would be acceptable in an anonymised dataset.

The number of groups which are smaller than threshold gives us a measure of how well a chosen combination of postcode component and population group anonymises the data. It is of course possible to aggregate small groups into larger ones, but this is best done when the majority of groups are already large enough to be anonymised.

2.2 Postcode Component Selection

We have used the area, district, sub-district and and sector components of the postcode; these represent increasingly smaller geographic areas, with the full postcode being the smallest. Where a postcode does not have a sub-district (the sub-district is mostly used in areas of London), we have substituted the district in its place.

In Fig. 1 we have plotted the percentage of of groups below the anonymity threshold, against the anonymity threshold for each postcode component and population group. Where the percentage of groups below the threshold is low, this means that the data would be well anonymised and vice versa. To use these plots:

1. Choose the population of interest.
    - For example, the total population ((a), top left graph)
2. Select an anonymity treshold on the x-axis.
    - This threshold should be chosen so that data is considered anonymised if all the geographic postcode groups are larger than the threshold
    - i.e. a treshold of 10^5 means that data is considered anonymous if the smallest group size is larger than 100,000.
3. The value of the curve at the chosen threshold tells us what percentage of groups are smaller than this threshold.
    - e.g. for a threshold of 10^5, only a small number the groups at the postcode area level (red curve) are smaller than 10,000 total population.
    - This would probably be considered anonymous at this level, as the remaining small groups can be aggregated into a single larger one.
    - In contrast, at the sector, sub-district of district level, close to 100% of the groups are below this threshold, and the data would be defined as not anonymous.

We see that each curve has a region where it rises very sharply. To the left of this steep region it is clear the data would be well anonymised, and to the right it would be poorly anonymised. Within the steep region it is less clear and it would be necessary to look at the small groups, how these might be aggregated and the impact that would have on use the data.

The largest viable threshold is found by looking a the largest postcode component and largest population group: “area” and “total population”. The curve (red line, top left of Fig. 1) begins to rise sharply at an anonymity threshold of approximately 100,000. This is explored further in Section 4.1.

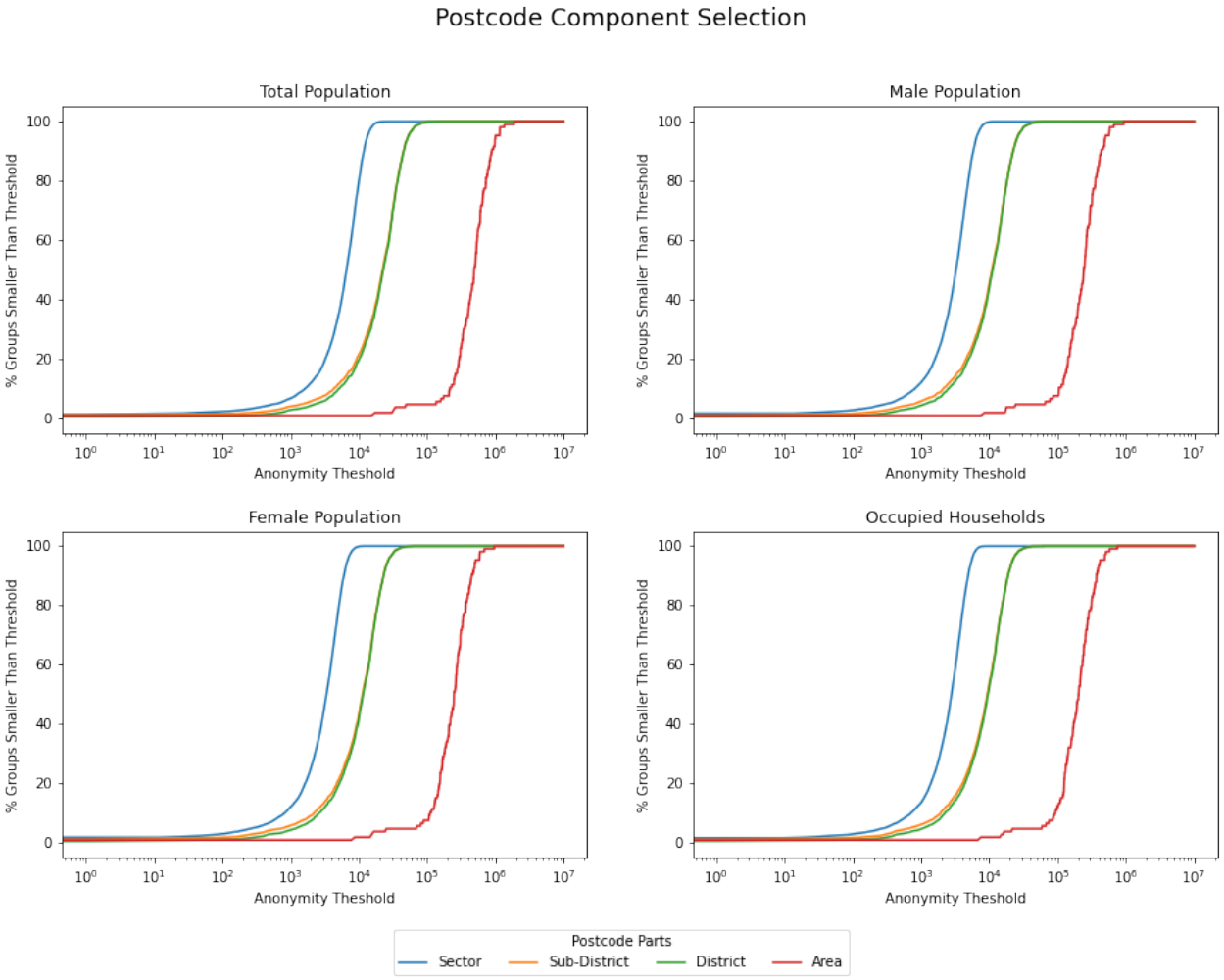

Figure 1: (a) top left: total population, (b) top right: male population, (c) bottom left: female population, (d) bottom right, occupied households.

3 Web App

A web app is available to assist selection. This is best used in conjunction with the above plots to first choose a region of interest, and then use the web app to explore the data in more detail.

4 Examples

4.1 Example A

- Population group: *Total Population*
- Anonymity threshold: $100,000$

Using Fig. 1a (top left), we see that all of the district, sub-district and sector group sizes are below the threshold, leaving only area. Using the lookup utility (Section 3) we find that there are 5 postcode areas with a population below the treshold, listed in Table 2. These five groups could be aggregated into one larger group that would be of size 138,322 - above the anonymity threshold.

Postcode Area	Total Population
DG	65
TD	18,331
EC	33,956
WC	35,745
LD	50,225

Table 2: Postcodes areas below the anonymity threshold for Example A

4.2 Example B

- Population group: *Occupied Households*
- Anonymity threshold: $1,200$

Area would be a confident choice in this case as only a single group is smaller than the selected anonymity threshold. However, if we want to maintain as much information as possible in the dataset, we may want to choose a smaller postcode component. At the sector level, 26.2% of groups are smaller than the anonymity threshold. Aggregating all of these into one group (of $\approx 770,000$ in size) would leave us with $\approx 6,000$ groups, many more than 105 groups we would have at the area level while still having an anonymised dataset. We would need to know more about the use case for the data before it is possible to recommend using "area" or "sector" (with aggregations).

5 Conclusions

We investigated if it was possible to use a similar system to that used in the USA for anonymising health data, where the zip-code is truncated to the three characters. We established that a simple truncation is not a good approach because of the way UK postcodes are constructed and we should instead use the different geographic components of the postcode to create groupings.

We defined an anonymity threshold: the smallest allowable group size for a dataset to be considered anonymised. Using a sample dataset, we created tools (graphs and web app) to assist in the selection of a postcode component for a given population group and anonymity threshold and were able to explore the impact on anonymisation of different combinations of these three factors.

Provided that we can set an anonymity threshold below $\approx 100,000$, then there will be some combination of postcode component and population group that, with some aggregation of small groups, would provide an anonymised data set.