US Style Postcode Privacy in the UK

Benjamin Collins

Introduction

In order to anonymize health data in the US, 5-digit Zip codes are rounded to 3-digits. This prevents small groups from being identified by their Zip code. These 3-digits correspond to the National area and the sectional centre or large city post office for that address with the final 2 removed digits being for the associate post office or delivery area. After this reduction, there are populations represented by the rounded Zip codes with less than 20,000 residents. These are merged together under a new code of 000 to prevent identifications.

The UK utilizes a similar postcode system with 5 to 7 digit postcodes. It is structured by having an outward code section, detailing the area and district, and an inward code section for the sector and unit.Applying the described method to UK postcodes for health data is possible and explored here.

Results

In the same way as is done in the US, the inward code was dropped, as each full postcode referenced an average of 43 individuals. Using outward codes alone, this average shifted to 24,000, over the threshold for Zip code merging used as standard in the US. However, a significant proportion of outwards codes fall under this 20,000 resident threshold as seen in Fig.1A.

Using the US method of creating a new outward code would create a code corresponding to 10 million individuals. The result would be a group 67 times larger than the biggest outwards code, corresponding to approximately 1/7th of the UK population. This would cause a significant decrease in locational health data information.

In order to retain as much information as possible without exposing individual data to privacy risk, it is suggested that smaller groups retain their area code and are merged into a larger district code of 00. Due to variation in the format of outward codes in the UK, this is not a simple step. Causing the most difficulty in this are outward codes of the format LETTER-NUMBER or LETTER-NUMBER-LETTER, e.g. N1 and N1C. Despite the different format these correspond to the same area. It important to ensure unrelated areas are not accidentality merged as this would corrupt the area information this method seeks to retain.

Taking the threshold standard of 20,000 residents reduces the number of codes below this threshold from 1010 to 12 and changes the distribution of total population in codes to that seen in Fig.1B. It also creates 25 new codes with a higher total residency than the previous maximum of outward codes. The residents of the codes below the threshold represent 0.31% of the UK population and are shown in Table 1. Additionally, of those

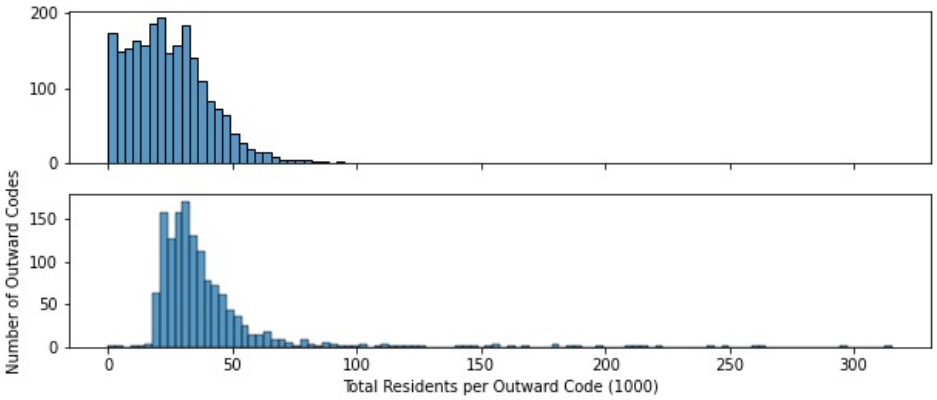

Figure 1: A set of histograms showing the distribution of total population (thousands) in outward codes, before (above or A) and after (below or B) the data anonymisation process described.

codes still below the threshold, 8 lie within only 25% of the 20,000 threshold, suggesting a different tolerance would enable more outward codes containing locational information.

Conclusion

From this investigation it is shown that a US style Zip code anonymisation is possible to use on UK postcodes. By reducing a UK postcode to its outward code and merging the area code of groups below a threshold size of 20,000, a majority of the UK population's location is anonymised with only 0.31% of the population having to be merged into a group without locational information. A proposed solution to this, as there only 12 cases, would be to merge them with an existing outward code in the same area where possible. Further investigation into total population threshold values may be able to yield a better balance between minimum size of total population in a postcode and proportion of population with loss of location based.

Table 1: A table listing the remaining 12 modified postcodes and the total population they represent.

Postcode	Total Population	Postcode	Total Population	Postcode	Total Population
DG00	65	PR00	15387	HG00	18374
SR00	5336	EN00	17203	SM00	18599
CR00	10220	SL00	17531	DH00	19570
UB00	14338	TD00	18331	WN00	19742