Technical Working Paper 8

Provision of data zones for SLS postcodes – a methodology

Kellas Campbell and Lee Williamson

Published online October 2022

Summary

This short technical working paper describes the methodology developed using NRS historical postcodes datasets to provision data zone for various SLS postcodes.

Contents

Summary 2

1 Background 4

2 Postcode and data zone issues 5

3 Methodology - working with the SPD 5

3.1 Overview of Postcode History data 6

3.2 Overview of methodology steps 6

3.2.1 Data Cleaning and Preparation 6

3.2.2 Dealing with Split Postcodes 7

3.2.3 Steps to retrieve the data zone 7

3.3 Miscellaneous 8

4 Conclusion 12

Appendix 1: SQL Query 13

References 14

1 Background

The Scottish Longitudinal Study (SLS) is a large-scale record linkage study. The SLS is a broad study containing many different types of data (Boyle *et al*, 2009). We currently link three censuses from 1991, 2001 and 2011, allowing over 20-year follow-up studies. We have also linked to other administrative data such as education data (school census and SQA) and the derived residential histories over a 15-year period from GP registrations (giving annual migratory data). We have extended the study back in time by creating a 1936 birth cohort, based on a linkage to the 1947 Scottish Mental Survey (a cognitive ability survey of almost all 11-year-olds in Scotland in 1947). This adds an important new dimension to the study allowing the examination of early life context (at 11) and later life outcomes if possible.

Most of the administrative datasets linked as part of the SLS have postcode information recorded, however, postcodes are restricted variables within the SLS, and researchers are not given access to postcodes together with payload data (i.e., the main research/cohort variables), in case of accidental disclosure if looking along the data row. Researchers work with postcodes more or less indirectly to protect privacy (ie postcodes can be used to add in other variables like datazones or ecological data which correspond to that postcode from look-up tables; the postcode can then be removed, before the researcher moves these new variables back to the payload dataset). Further, for most projects there is no research interest in postcodes but rather in small area geography such as Scottish Neighbourhood Statistics (SNS) data zones. The SNS data zones likewise are considered restricted variables within the SLS given there are just under 7000 for both 2001 and 2011 (ie different data zones). However, researchers are allowed to access these data zones from the SLS Safe Setting to undertake analysis (i.e., using them for multilevel modelling or mapping on SIMD or sub-domains for their year of interest).

For the 2001 and 2011 Censuses we have as part of the Census linkage – provided from NRS – a look-up from Census postcode to data zone based on ‘Output Area’ geography mapping. However, we were not provided such a look-up for address or migrant address one year ago. We also have postcodes from the Vital Events data (i.e., births, deaths, marriages), Education data (up to 2014), and the annual migratory data (i.e., GP histories up to 2018), all without data zones.

Within recent years we have had several SLS research projects requesting derived data zones for these four sources of postcodes listed above, for which we do not have data zones; as such we have developed a methodology to provide them. We are not planning to add these data zones to the SLS database as new variables, as they are not an SLS product or derived variable. However, SLS staff can apply this methodology to a postcode of interest and provide the resulting data zones for the researcher. Further, to avoid confusion with other data zones held within the SLS (which have been provided by NRS), we are not planning to hold all the possible data zones within the SLS database.

2 Postcode and data zone issues

Postcodes can change over time, and in some cases they are deleted, and then reintroduced/reissued. When this is due to the demolition and/or redevelopment of housing, Royal Mail will usually wait two or three years before reassigning the postcode of the demolished block to a new location. Sometimes, though, a postcode is still active, but there is an input error to be fixed, or more accurate grid coordinates available; so in these situations, the postcode will be deleted and reintroduced within only a few months — sometimes even on the same day.

Thus, some postcodes have more than one record in the SPD (Scottish Postcode Directory) datasets. For example, in NRS's 2021-02 release of *SmallUser.csv,* KA11 5AR appears the most often with six active periods. Two postcodes appear five times, 19 postcodes appear four times and 321 postcodes appear three times; the remaining 176,634 postcodes appear only once. These numbers are after removing postcodes flagged as *NeverDigitised* and *B* and *C* split postcodes, as described below.

There was a substantial change to postcodes in the 1990s which makes linking SNS data zones to – for example – 1991 Census postcodes difficult. Our NRS colleagues provided us with a 1991 Census postcodes lookup file to both 2001 and 2011 data zones. However, this does not help with any Vital Events postcodes from the 1990s or the Census postcode from address one year ago.

Further, the format of postcodes – depending on the source – can differ. For example, some can have spaces in the middle and some do not, meaning some pre-processing is required before applying the methodology outlined here to return the data zones.

3 Methodology - working with the SPD

The methodology uses the Scottish Postcode Directory (SPD) produced by NRS Geography, specifically, the 2021-2 release of the Postcode Index, the source data being *SmallUser.csv* and *LargeUser.csv*.

LargeUser.csv comprises postcodes receiving 1,000 or more items a day (e.g., a business); for example, a place of birth variable, such as *rbplapc*, will in many cases match a record from *LargeUser*, due to it being the postcode of a hospital.

3.1 Overview of Postcode History data

The SPD Postcode Index extract provides full postcode history, including deleted and reintroduced postcodes. Looking at the example below (*table 1*), AB11 5FA has had three active periods. An event date of 10 May 1999 would return row 3 (i.e., *DataZone2011Code* S01006589). An event date of 02 January 2009 would return record 2; while it falls within the period of record 3, it also falls within the period of record 2, since we apply a ten-month buffer to account for administrative lag, and our methodology always returns the most recently updated record in these cases.

	Postcode	DateOfIntroduction	DateOfDeletion	DataZone2011Code	Latitude	Longitude
1	**AB11 5FA**	2013-10-17	<NA>	S01006637	57.15015	-2.086840
2	**AB11 5FA**	2009-04-20	2009-10-14	S01006637	57.14878	-2.087696
3	**AB11 5FA**	1996-04-01	2009-01-05	S01006589	57.14819	-2.091694

Table 1 AB11 5FA from SPD’s SmallUser.csv (2021-2)

3.2 Overview of methodology steps

3.2.1 Data Cleaning and Preparation

Upon import into the lookup database, we process the *SmallUser.csv* and *LargeUser.csv* in the following manner:

1. Because some postcode variables may have an extra space or no space at all, whitespace is stripped.
2. Administrative postcodes that are indicated by *NeverDigitised* = ‘Y’, are removed. These postcodes do not have a geography.
3. Split postcodes are indicated by *SplitIndicator* = ‘Y’; in these cases, the ‘A’ record is retained, and the others removed.
4. The table is sorted by Postcode and then by *DateOfIntroduction* in descending order, so that newest records come first.

3.2.2 Dealing with Split Postcodes

When postcodes are split over one or more data zones, NRS Geography appends them with ‘A’, ‘B’ or ‘C’ and sets *SplitIndicator* to ‘Y’. ‘A’ represents the most populous part of the postcode and thus we use these for looking up the data zone. In the 2021-2 release of *SmallUser.csv*, there are 1,243 records flagged as split. Of these, 621 are ‘A’.

Shown below is R code used for retaining only the ‘A’ split postcode record:

```
data2 <- data[(SplitIndicator=="Y" & grepl("A$", Postcode)) | SplitIndicator=="N"]
data2[SplitIndicator=="Y",Postcode f:= gsub('.{1}$','',Postcode)]
```

To illustrate the above, let us take the postcode AB1 0LT. For this example, we will say this is a 1991 census postcode, from the variable *enumpc*9, given that we need a time element. So retaining only the ‘A’ split postcodes, this would return S01006857 (from row 1, see table 2 below); and this would be the 2011 data zone, *datazone1*:

	Postcode	DateOfIntroduction	DataZone2011Code	SplitIndicator
1	AB1 0LTA	1973-08-01	S01006857	Y
2	AB1 0LTB	1973-08-01	S01006506	Y

Table 2 Split postcode example

3.2.3 Steps to retrieve the data zone

When producing a project data extract, here are the steps taken to return the SNS data zones:

1. Search the SPD postcode records looking for a match on *Postcode*.
2. If the event date falls between the matched record’s *DateOfIntroduction* minus ten months and *DateOfDeletion* plus ten months, or if it falls after *DateOfIntroduction* minus ten months and *DateOfDeletion* is null — indicating an active postcode — return the record.
3. If there is more than one returned record, return the newest record according to *DateOfIntroduction*.

We chose to use a buffer of ten months on either side of a postcode's active period to account for administrative lag and the fact that the NRS updates this information on a quarterly basis. The number of months settled upon after experimenting with other thresholds. However, if there is interest in a lower threshold of eight months, the approach can be modified.

Because some postcodes, such as the Vital Events postcode on place of birth (*rbplapc),* could represent a non-residence such as a hospital, we also check for it in the LargeUser table, following the same method.

The code, which runs in t-SQL, is given in Appendix 1.

3.3 Miscellaneous

To understand the impact of postcodes potentially moving data zones, a few investigations were undertaken in R, and resulting the charts are shown below. This is from the data in *SmallUser.csv*. We added a Boolean variable, *datazone_same*, which tells us whether a postcode's data zone changed upon reintroduction:

```
data[,datazone_same := (DataZone2011Code !=
shift(DataZone2011Code,1,type="lead")), by="Postcode"]
```

Investigating the cases where the data zone changed after the postcode was reintroduced:

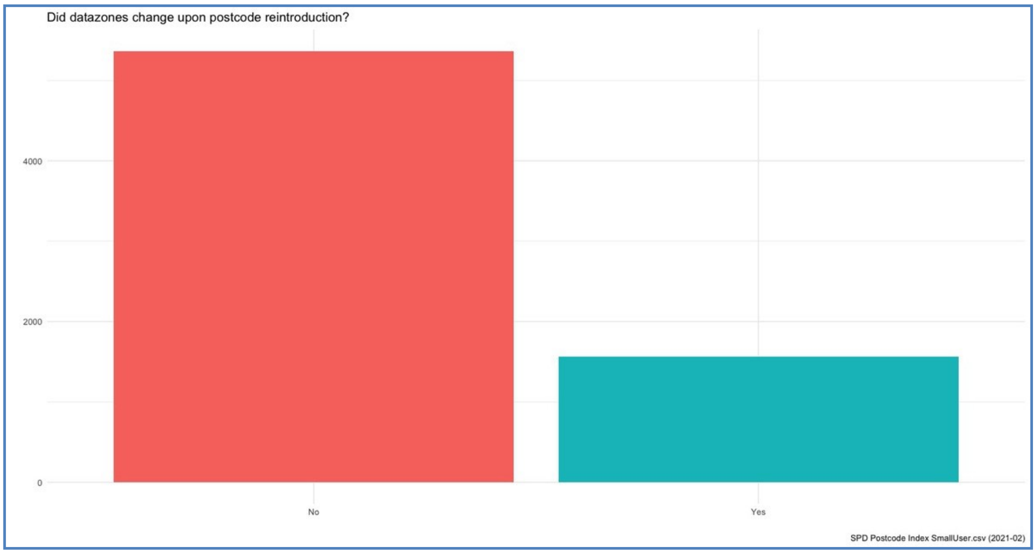

Figure 1 Data zone changes upon postcode reintroduction

Investigating the distance change (in meters) for postcodes that were reintroduced, by the time to reuse (in days). Figure 2 shows this relationship with the number of days split into 10 bins, as described in the table below.

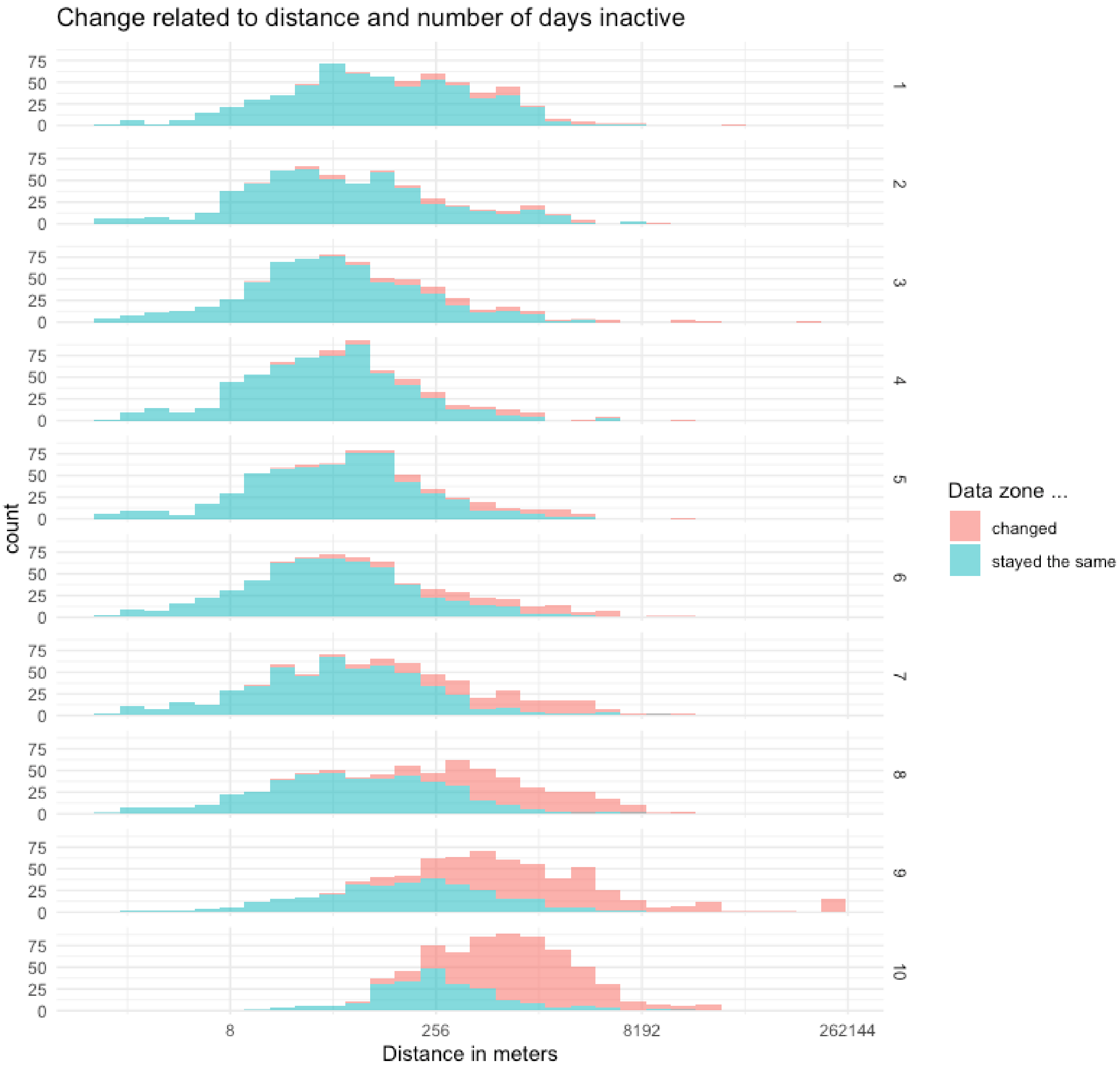

Figure 2 Data zone change ~ distance ~ by the number of days before postcode reintroduction

(by 10 bins – ie not 1 to 10 days, please see table 3 below for details on days per bin)

Days inactive bin number	mean	median	min	max	sd
1	0.00000	0.0	0	0	0.00000
2	59.94228	62.0	0	92	32.19332
3	156.44733	163.0	92	255	42.02321
4	338.47619	335.0	257	438	56.01823
5	569.50072	572.0	438	725	94.17943
6	944.20058	946.0	725	1188	134.86322
7	1636.98124	1629.0	1188	2180	279.11621
8	3071.17172	2997.0	2180	4352	615.25075
9	6146.07504	6217.0	4355	8118	1076.91454
10	11525.43931	11179.5	8118	16402	2247.92387

Table 3 Bin values for number of days a reintroduced postcode was inactive in figure 2

Investigating the cases where the data zone changed after the postcode was reintroduced, by length of time for reintroduction (in days):

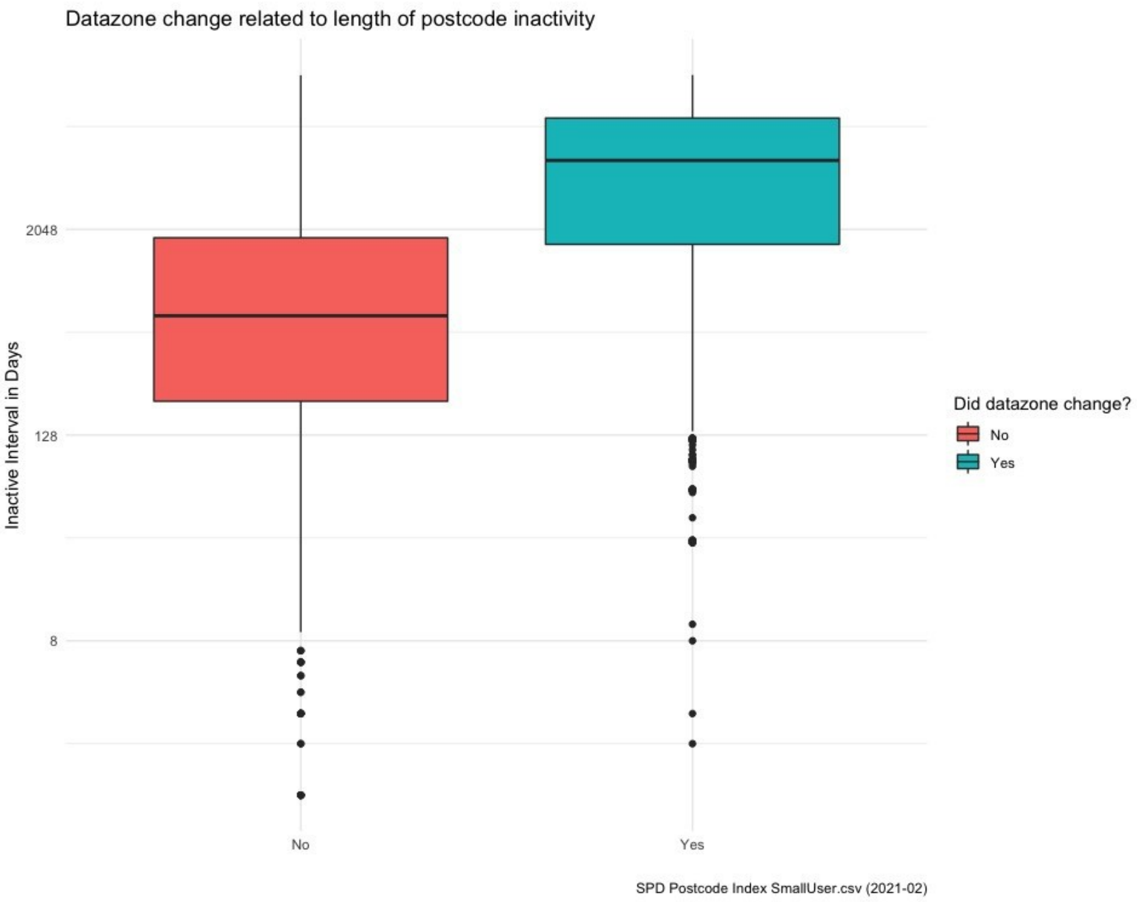

Figure 3 Data zone change related to postcode inactivity

4 Conclusion

This paper presents a pragmatic methodological approach of providing SNS data zones (both 2001 and 2011) to the SLS postcode variables which do not already have a data zone look-up provided by NRS. These variables without data zones look-ups from NRS, are from the four data sources: address one year ago, the Vital Events data (i.e., births, deaths, marriages), Education data and the annual migratory data (i.e., from GP registrations).

The provision of data zones from postcodes for these data sources, allows SLS researchers to work directly with their data extracts. Since postcodes are considered restricted level 2 variables, but data zones are restricted level 3 variables - so researchers can work with data zones within the SLS Safe Setting for analysis (i.e., for multilevel modelling or adding in SIMD/domain scores). However, no output can be produced/cleared to take out the SLS Safe Setting at data zone level.

This methodology presented here provides a way for SLS researchers to access the geography variables required for their research, and has been developed in direct response to SLS researcher needs. However, researchers are still free to bring in their own look-up table(s), which can be linked on to postcodes to provide data zones, if their own method is preferred. Further as noted, we are not planning on holding all these derived data zones as part of the main SLS database, as all the other data zones were provided by NRS at source (or by NRS look-ups).

Lastly, this paper does not cover the more theoretical questions over providing data zones—which are a statistical geography based on a population size (of between 500-1000 at either 2001 or 2011)—or other time periods for which the population size does not relate to.

Appendix 1: SQL Query

```
SELECT TOP 1 WITH TIES

v.slsno, v.pc_start_date, p.postcode, p.datazone2001code, p.datazone2011code
FROM v

JOIN

slslookups.dbo.tblsls21_L_PC_INDEX_SMALL p ON (REPLACE(v.postcode,' ','') =
REPLACE(p.postcode,' ',''))

WHERE A

v.pc_start_date > dateadd(month, -10, dateofintroduction) AND (DateOfDeletion
IS NULL OR dateadd(month, 10, dateofdeletion) >= v.pc_start_date)
```

References

1. Paul J Boyle, Peteke Feijten, Zhiqiang Feng, Lin Hattersley, Zengyi Huang, Joan Nolan, Gillian Raab, Cohort Profile: The Scottish Longitudinal Study (SLS), *International Journal of Epidemiology*, Volume 38, Issue 2, April 2009, Pages 385–392, https://doi.org/10.1093/ije/dyn087
2. https://www.scotlandscensus.gov.uk/about/2011-census/2011-census-geographies/
3. 2021-2 Scottish Postcode Directory Files | National Records of Scotland https://www.nrscotland.gov.uk/statistics-and-data/geography/our-products/scottish-postcode-directory/2021-2
4. Geography – Background Information – Postcodes. History of Digitizing Postcodes in Scotland https://www.nrscotland.gov.uk/files/geography/Products/postcode-bkgrd-info.pdf
5. Scottish Index of Multiple Deprivation 2020 https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/?utm_source=redirect&utm_medium=shorturl&utm_campaign=simd