Notes on Scottish Postcode Directory Quality Checking

Gerald Leung

Public Health Scotland

Draft version: March 4, 2022

Contents

1 Introduction

In this short document, we summarise the process of the Scottish Postcode Directory (SPD) file checks from the National Records of Scotland (NRS) by the Geospatial Team at Public Health Scotland (PHS). In short, NRS updates SPD files twice a year, around March and September[1]. The Geospatial Team will then conduct quality checks of the updated files and compare any changes with the previous version. They will then produce lookup files in various formats, such as `.csv`. This document focuses on summarising the processes of quality checking (QC) and lookup files production using R. Here we also reproduce a brief summary of the data, which can be found in Appendix A. Full details can be found through the data dictionary provided by the NRS, or accessed through

```
//PHIBCS/PHI/Referencing & Standards/GPD/1_Geography
/Scottish Postcode Directory/Source Data/2021_2
/ISD Data Dictionary_2021-2
```

for the **2021-2** version.

[1] https://www.nrscotland.gov.uk/statistics-and-data/geography/nrs-postcode-extract

2 Initial Quality Checks

The R scripts are located in

```
//PHIBCS/PHI/Referencing & Standards/GPD/5_GitHub/GPD/Geography
/Scottish Postcode Directory
```

and the Standard Operating Procedures (SOPs) are located in

```
//PHIBCS/PHI/Referencing & Standards/GPD/1_Geography
/Scottish Postcode Directory/SOPs.
```

The source files can be found in

```
//PHIBCS/PHI/Referencing & Standards/GPD/1_Geography
/Scottish Postcode Directory/Source Data.
```

They can be accessed through the organisation's VPN. The first QC is done through the R script `1_Check NRS SPD.R`.

In short, this script conducts a general initial QC of the newest version of SPD files and compares changes with the previous version. Figure 1 shows roughly the structure of the script and the QC process.

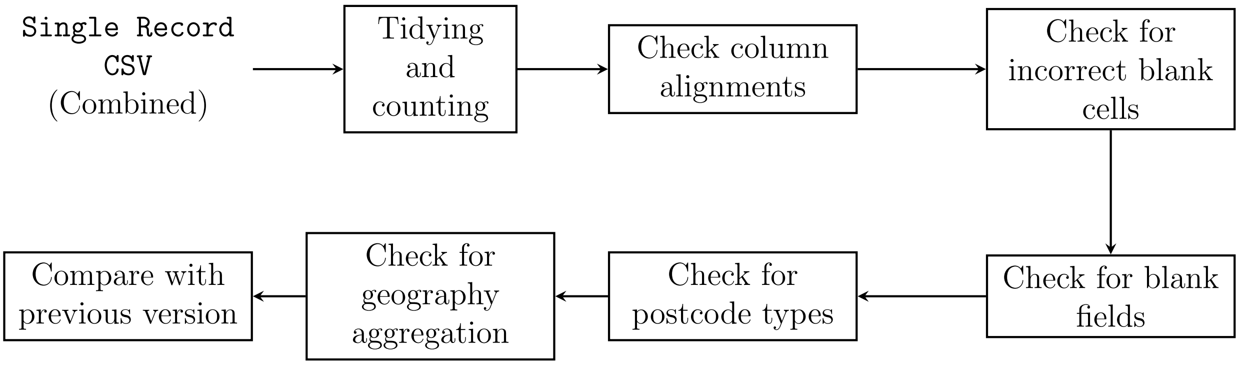
Single Record CSV (Combined)
Tidying and counting
Check column alignments
Check for incorrect blank cells
Check for blank fields
Check for postcode types
Check for geography aggregation
Compare with previous version

Figure 1: Schematic diagram of the initial QC process.

To start with, all empty rows contained in the files are removed and the files are combined. The user will then record the total number of records and postcode types (`Large User` or `Small User`). Sanity checks are conducted to make sure there are no blank entries for Health Board and Council Areas. The script then checks for postcodes that moved from Glasgow City Council to North Lanarkshire Council (eight in total). Checks are then carried out to make sure columns are aligned correctly. For example, between Health Board Area Code and Council Area Code columns to make sure their values in 1995 (e.g. `HealthBoardArea1995Code`) represent the same area in 2006 (e.g. `HealthBoardArea2006Code`). This is generally done by assigning a number to the case when a particular column does not have an expected code that should match with the other column. For instance:

```
HB_issue = case_when(HealthBoardArea1995Code == '01' &
                              HealthBoardArea2006Code != 'S08000008'
    ~ 1)
Count(HB_issue)
```

Refer to Appendix A or the NRS data dictionary mentioned in Section 1 for more information of the data.

Extensive checks are then carried out to make sure there are no blank fields, through the `checks` function

```r
checks <- function(variable1, variable2){

  spd %>%
    mutate(check = case_when(variable1 == "" & variable2 != "" ~ 1,
                             variable1 != "" & variable2 == "" ~ 2))
    %>%
    count(check) %>%
    print()

}
```

So for instance, we should expect that if `OutputArea2001Code` is blank, then `DataZone2001Code` should also be blank since the data zones are made up of output areas. Throughout the script, blank fields are also checked by, for example,

```r
spd %>% summarise(missing = sum(is.na(DataZone2001Code)))
```

where `spd` refers to the `SingleRecord.csv` data frame.

Examinations are also carried out to make sure there are no English voting codes included. Some checks are also carried to make sure there are no errors with postcodes and their split indicators. Finally, the script makes sure that mappings are done correctly, such as Data Zones and Health Boards.

The script also makes comparisons with previous SPD files. For instance,it checks for changes in postcodes (some have been deleted), different data zones and health boards etc.

3 Lookup Files

The next step of the process is to produce lookup files for other uses and analysis within the organisation. This is done through the `2_Create Postcode Lookup Files.R` script. Figure 2 shows an overview of the process. To begin with, all the variables are renamed and changed into correct format. The Urban Rural codes are changed into named columns. Some formatting are done to ensure consistency with R and SPSS. For instance, by adding leading zeros to some columns and making sure blank cells are represented as `NA`. The script then connects to the PHS Open Data website[2] to access Geography Codes and Names data file, which are then used for column names, including Data Zone, Intermediate Zone, Council Area, Health and Social Care Partnership and Health Board. These are then merged with the original `SingleRecord.csv` to produce the lookup file.

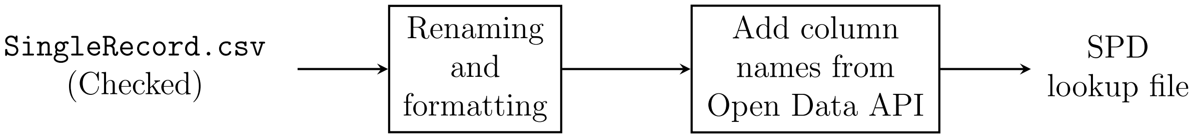

Figure 2: Schematic diagram of the lookup file production process.

[2] https://www.opendata.nhs.scot/

4 Updating Postcode Table

After QC, the data needs to be uploaded to the corporate data warehouse (CDW). The lookup files are found in

```
//PHIBCS/PHI/Referencing & Standards/GPD/1_Geography
/Scottish Postcode Directory/Lookup Files.
```

The tables from the previous SPD files have to be updated to satisfy any requirements from the IT. Some columns are not required and therefore removed. On the other hand, there are columns that are no longer supplied by the NRS. They are however still required for CDW and therefore must be added here. The columns are renamed and finally reordered, and saved as

```
//PHIBCS/PHI/Referencing & Standards/GPD/1_Geography
/Scottish Postcode Directory/Lookup Files
/Scottish_Postcode_Directory_2021-2_CDW.csv
```

for the **2021-2** version.

5 Testing Postcode Table

The file is then uploaded to APXU by IT, which requires testing. This is done with the script `5_Check Postcode table on MGSREF APXU`. We start off by connecting to the APXU and access the database there.

A Appendix: Data Dictionary

Here we reproduced and summarised the data dictionary provided by the NRS. Table 1 shows the rough definition of the data within `SingleRecord.csv` in order.

Name	**Type**	**Notes**
CensusHouseholdCount1991	Integer	1991 Census postcode usually resident household count.
CensusHouseholdCount2001	Integer	2001 Census postcode household count.
CensusHouseholdCount2011	Integer	2011 Census postcode household count.
CensusPopulationCount1991	Integer	1991 Census postcode residential count.
CensusPopulationCount2001	Integer	2001 Census postcode residential count.
CensusPopulationCount2011	Integer	2011 Census postcode residential count.
CivilParish1930Code	Character	1930 Civil Parish code.
CommunityHealthPartnership2004Code	Character	2004 Community Health Partnership code ***(ISD version only)***.
CommunityHealthPartnership2007Code	Character	2007 Community Health Partnership code ***(ISD version only)***.
CommunityHealthPartnership2011Code	Character	2011 Community Health Partnership code ***(ISD version only)***.
CommunityHealthPartnership2012Code	Character	2012 Community Health Partnership code ***(ISD version only)***.
CommunityHealthPartnershipSubAreas2011Code	Character	2011 Community Health Partnership sub sector code ***(ISD version only)***.
CouncilArea2011Code	Character	2011 Council Area code ***(ISD version only)***.
CouncilArea2018Code	Character	2018 Council Area code ***(ISD version only)***.
CouncilArea2019Code	Character	2019 Council Area code.
DataZone2001Code	Character	2001 Data Zone built from 2001 Output Areas.

DataZone2011Code	Character	2011 Data Zone built from 2011 Output Areas.
DateOfIntroduction	Date	Postcode introduction date. Currently shown as #### on the `csv` file.
DateOfDeletion	Date	Postcode deletion date. Similarly to date of introduction.
DeliveryPointCount	Integer	The Royal Mail delivery point count.
DeliveryPointCountNonResidential	Integer	Non-residential delivery point count.
ElectoralWard2019Code	Character	2019 Electoral Ward code.
EnterpriseRegion2008Code	Character	2008 Scottish Enterprise Region code.
GridLinkIndicator	Character	Whether grid reference has been assigned via GridLink (Y/N).
GridLinkPositionalAccuracy	Character	Positional accuracy of the GridLink grid reference allocated to each postcode. See full dictionary for details.
GridReferenceEasting	Character	Grid reference easting.
GridReferenceNorthing	Character	Grid reference northing.
HealthBoardArea1995Code	Character	1995 Health Board code.
HealthBoardArea2006Code	Character	2006 Health Board code.
HealthBoardArea2014Code	Character	2014 Health Board code ***(ISD version only)***.
HealthBoardArea2018Code	Character	2018 Health Board code ***(ISD version only)***.
HealthBoardArea2019Code	Character	2019 Health Board code.
HouseholdCount	Integer	The Royal Mail Household count.
Imputed	Character	Whether postcode's higher area values have been inputed ***(ISD version only)***.

IntegrationAuthority2016Code	Character	2016 Integration Authority code (Health and Social Care Partnership) ***(ISD version only)***.
IntegrationAuthority2018Code	Character	2018 Integration Authority code (Health and Social Care Partnership) ***(ISD version only)***.
IntegrationAuthority2019Code	Character	2019 Health and Social Care Partnership code.
IntermediateZone2001Code	Character	2001 Intermediate Zone built from 2001 Data Zones.
IntermediateZone2011Code	Character	2011 Intermediate Zone built from 2011 Data Zones.
Islands2021Code	Character	Island identification code.
Latitude	Numeric	Coordinates in degrees.
Longitude	Numeric	Coordinates in degrees.
LAU2019Level1Code	Character	European Area Classification Local Administrative Unit (LAU) 2019 level 1. E.g. Council areas.
LinkedSmallUserPostcode	Character	Small User postcode that contains the grid reference of the Large User postcode.
LinkedSmallUserPostcodeSplitChar	Character	Split indicator for small user, A, B, C ***(ISD version only)***.
Locality1991Code	Character	1991 Locality code.
Locality2001Code	Character	2001 Locality code.
Locality2016Code	Character	2016 Locality code.
LocalGovernmentDistrict1991Code	Character	1991 Local Government District code.
LocalGovernmentDistrict1995Code	Character	1995 Local Government District code.
NationalPark2010Code	Character	2010 National Park code.
NeverDigitised	Character	Whether a postcode boundary has been digitised (Y/N).

NUTS2018Level2Code	Character	European Area Classification Nomenclature of Territorial Units for Statistics (NUTS) 2018 level 2.
NUTS2018Level3Code	Character	European Area Classification Nomenclature of Territorial Units for Statistics (NUTS) 2018 level 3.
OutputArea1991Code	Character	1991 Census Output Area code.
OutputArea2001Code	Character	2001 Census Output Area code.
OutputArea2011Code	Character	2011 Census Output Area code.
Postcode	Character	The Royal Mail postcode. Consists of area, district, sector and unit. E.g. AB1 0AA.
PostcodeDistrict	Character	Postcode District. E.g. AB1.
PostcodeSector	Character	Postcode Sector. E.g. AB1 0.
PostcodeType	Character	Small (S) or Large (L) ***(ISD version only)***.
RegistrationDistrict2007Code	Character	2019 Council Area (same as Council Area boundaries).
ROACommunityPlanningPartnership2006Code	Character	2006 Regeneration Outcome Area Community Planning Partnerships (CPP) code.
ROALocal2006Code	Character	2006 Regeneration Outcome Area Local code.
Settlement2001Code	Character	2001 Settlement code.
Settlement2016Code	Character	2016 Settlement code.
ScottishIndexOfMultipleDeprivation2020Rank	Character	Rank Data Zones (2011) according to deprivation.
ScottishParliamentaryRegion2021Code	Character	2021 Scottish Parliamentary Region code.
ScottishParliamentaryConstituency2021Code	Character	2021 Scottish Parliamentary Constituency code.

SplitChar	Character	Split indicator A, B, C ***(ISD version only)***.
SplitIndicator	Character	Split postcode indicator. Y/N.
StrategicDevelopmentPlanningArea2013Code	Character	2013 Strategic Development Planning Area code.
TravelToWorkArea2011Code	Character	2011 Travel to Work Area code.
UKParliamentaryConstituency2005Code	Character	2005 UK Parliamentary Constituency code.
UrbanRural2Fold2016Code	Character	Standard definition of rural Scotland following the Scottish Government Urban Rural classification (2 fold) ***(ISD version only)***.
UrbanRural3Fold2016Code	Character	Similarly to above (3 fold) ***(ISD version only)***.
UrbanRural6Fold2016Code	Character	Standard definition of rural Scotland following the Scottish Government Urban Rural classification (6 fold).
UrbanRural8Fold2016Code	Character	Similarly to above (8 fold).

Table 1: Data dictionary of the SPD file. Reproduced from NRS.