**R Script review – PEoLC publication**

Written by: Ryan Harris 2021

Last Update: 07/12/2023 By: Sissi Pizzo

00\_geospatial-package-install.R

The Geospatial R Packages has been included in the 2022/23 annual publication codes because of the move from R server & R desktop version to Posit.

This package is available in GitHub and the [PHS Data Science - Knowledge Base (public-health-scotland.github.io)](https://public-health-scotland.github.io/knowledge-base/docs/Posit%20Infrastructure?doc=How%20to%20Install%20and%20Use%20Geospatial%20R%20Packages.md) but you must follow the directions from the PHS website of Knowledge Base because is here where the Data Science Team add the latest information.

If you run the script for the very first time, you must run it in blocks:

* Lines 29-45 (First Block)
* Lines 48-57
* Lines 62-66
* Lines 67-77
* Lines 79-82
* Lines 85-129
* Lines 131-132 (Last Block)

IMPORTANT: if all the packages are installed, then you must run only the First and the Last Block and skip all the other part.

00\_setup-environment

* Lines 19-46, these load the packages required for running this publication. If any of these don’t run when selected make sure they are installed using install.packages(“”).
* Lines 66-90 will be used to populate text where dates are needed. These lines need to be updated every year. It also creates the date format in order to display correctly in outputs and also to read in correct data from SMR using required time periods.
* Lines 139-217, reads in lookup files (postcode, SIMD and locality) keeping only the relevant variables within each file and also renames certain variables for future matching.

01\_create-basefile

* Lines 32-69, reads in GRO deaths, SMR01 (including GLS) and SMR04 using dates defined in previous syntax file. (Access to these data sets are required via data authoriser in order to be able to run)
* Lines 85-112, these are just simple aggregate functions for the SMR01 and SMR04 data marts separately, breaking by link number. Here the data is aggregated to CIS level, which includes community care home episodes, these care home episodes are then flagged and removed as they shouldn’t be counted as part of SMR activity.
* Lines 118-133, this section combines the SMR01 and SMR04 files, an aggregate function is then used to remove records where stays overlap from both datasets, again breaking by link number.
* Lines 138-170, this section calculates the date of 6 months before death and then the length of stay. Firstly, remove 183 days from the date of death to create the new ‘six months’ variable (140-141). Next, we modify and then select out the admission date to ensure only stays within the last 6 months of life are selected, admission dates after date of death are also excluded (152-153). Next, modify the discharge date, where the discharge date is greater than date of death, make equal to date of death (156-158). Now we calculate the length of stay using the time interval between the modified admission and discharge dates (161-162). Finally, aggregate by patient level and accumulate the length of stay for each patient and then modify the length of stay if 183 to 182.5, which is exactly 6 months (165-170).
* Lines 176-180, match on postcode, SIMD and locality information to the deaths data set.
* Lines 186-197, match on deaths data set to SMR by link number, keeps required variables, and aggregates data.
* Lines 205-210 read in open data for data completeness, however, depending on when this is run, the table may not be up to date. For this, a workaround is available (line 215) – this should be run ONLY if 205-210 doesn’t work.

IMPORTANT: 30/11/2023 The completeness function should be produced late June/beginning of July this because of how the SMR completeness by quarters are published: ‘previous 6 quarters for New and Return SMR00 and SMR01 Acute only previous 4 quarters for SMR02, SMR04 and SMR01 Geriatric Long Stay’. So to match the latest 4 quarters of the publication for SMR01 GLS and SMR04, outputs need to be produce in June. When running the code in August/September for the final publication remember to adjust lines 715-752 of Summary.Rmd.

02\_old-method

* Lines 25-27, Read in SMR01, SMR04 and deaths data files.
* Lines 34-48, aggregate SMR01 and SMR04 data by link number and CIS level. Again during this aggregation, community care home episodes are included initially, flagged then excluded in order to match the process of the previous script.
* Lines 54-69, combine SMR01 and SMR04 data and aggregate by link number where stays overlap in order to remove these.
* Lines 74-106, this section calculates the date of 6 months before death and then the length of stay. Firstly, remove 183 days from the date of death to create the new ‘six months’ variable (74-77). Next, we modify and then select out the admission date to ensure only stays within the last 6 months of life are selected, admission dates after date of death are also excluded (80-89). Next, modify the discharge date, where the discharge date is greater than date of death, make equal to date of death (92-94). Now we calculate the length of stay using the time interval between the modified admission and discharge dates (97-98). Finally, aggregate by patient level and accumulate the length of stay for each patient and then modify the length of stay if 183 to 182.5, which is exactly 6 months (101-103).
* Lines 112-116, match on postcode, SIMD and locality information to the deaths data set.
* Lines 122-137, create final file with ‘los\_old’ variable and attach to end of base file created at the end of the previous script.

03\_create-figures

* Lines 24-31, reads in base file from previous scripts and modify financial year variable if this is a provisional publication.
* Lines 38-75, this creates figure 1. Firstly, the data source to be used is selected and all years of interest are to be used, with the measures of data being percentage spent in hospital or at home/community (38-45). Next, the order of the stacked bar chart is defined and the design of the graph is set up (54-57). The next section formats the design of the graph, such as font size of axis titles and legend (58-70). Finally, the axis labels are defined (68-70). The graph is then saved into the specific folder, in this case two different versions are saved with varying height and width (75-84).
* Lines 86-219, this section creates the health board map. The data file is set up to create the map, using basefile again and marking the health board as the measure (86-100). Next the design of the map is created, using geom\_polygon and latitude and longitude as the axis design, axis titles and text aren’t required for the map so are kept blank, with some text to mark the health boards with the lowest and highest percentages (102-140). Letters to identify each NHS board are added to the map (143-212). Then the map is then saved in the required folder (216-219).
* Lines 223-259, creates the age and gender chart. Firstly, some modification is required to basefile in order to define ‘All Ages’ as well as the original age groups and add on to the basefile (229-234). Next the format of the clustered bar chart is created, with the x axis being age group, y axis being the measure, which is percentage and the filled bars being patient gender, and defining axis headings (238-252). The graph is then saved in the markdown folder (256-259).
* Lines 262-290, creates the deprivation chart. Firstly, define that the basefile is the data set to use, with SIMD being the variable of interest (267-268). The next section creates the chart, which in this case is just a standard bar chart with SIMD as the x axis and percentage on the y axis, labelling the axes with specific font design and size as well (272-284). Then the chart is saved in the markdown folder (288-291).
* Lines 294-325, this creates the urban/rural chart. Firstly, define that the basefile is the data set to use, with urban\_rural being the variable of interest (299-302). The next section creates the chart, which in this case is just a standard bar chart with urban/rural classification as the x axis and percentage on the y axis, labelling the axes with specific font design and size as well (306-318). Then the chart is saved in the markdown folder (322-325).
* Lines 328-382, creates multiple line charts by health board and Scotland. Firstly, we need to modify the basefile to include a Scotland total as an additional health board in order to display alongside all other boards (332-335). The next section designs the line charts with financial year of death along the x axis and the measure, in this case percentage as the y axis variable, along with axes labels and font design (351-367). Then each graph is given a heading from Scotland to start, then each health board in alphabetical order placed in a specific position on each chart (368-373). Then the final chart is saved in the markdown folder (377-382).
* Lines 385-429, creates multiple line charts by health and social care partnership. Firstly, we select the basefile as the data set to be used (388-391). The next section designs the line charts with financial year of death along the x axis and the measure, in this case percentage as the y axis variable, along with axes labels and font design (397-413). Then each graph is given a heading for each HSCP in alphabetical order placed in a specific position on each chart (415-419). Then the final chart is saved in the markdown folder (426-429).
* Lines 432-485, creates multiple line charts, one for each deprivation quintile. Firstly, we select the basefile as the data set to be used and rename the deprivation quintiles to ensure they are more readable when displayed on the charts (436-446). The next section designs the line charts with financial year of death along the x axis and the measure, in this case percentage as the y axis variable, along with axes labels and font design (452-467). Then each graph is given a heading for each deprivation quintile from 1 – most deprived to 5 – least deprived in a specific position on each chart (468-478). Then the final chart is saved in the markdown folder (480-485).
* Lines 488-533, creates multiple line charts, one for each urban/rural description. Firstly, we select the basefile as the data set to be used (491-494). The next section designs the line charts with financial year of death along the x axis and the measure, in this case percentage as the y axis variable, along with axes labels and font design (500-514). Then each graph is given a heading for each deprivation quintile from 1 – large urban areas to 6 – remote rural in a specific position on each chart (515-525). Then the final chart is saved in the markdown folder (529-533).
* Lines 535-589, creates a final line chart which compares the old measure with the new measure which excludes care home activity. Firstly, we need to confirm the basefile as the data set and also redefine the measures in order to distinguish between old and new, these are then renamed as follows, qom\_old = Old measure and qom\_new = New measure (excluding Care Home activity) (539-560). The next section designs the line chart with financial year of death along the x axis and the measure, in this case percentage as the y axis variable, along with axes labels and font design, with the old measure being marked with a dotted line and the new measure with a filled line (561-582). Then the final chart is saved in the markdown folder (586-589).

04\_create-excel-tables

* Lines 24-30, creates link to the main report on PHS website.
* Lines 34-35, reads in basefile as data set to be used for creating excel tables.
* Lines 45-144, restructures basefile into multiple different excel outputs to populate the qom table. This creates an excel file adding different rows by category, with a category split and the measure, which in this case is the percentage of last six months of life spent at home or in a community setting.
* Lines 49-54, creates the Scotland (category) level outputs for all financial years.
* Lines 56-61, creates the health board (category) output for all financial years with health board name the category split.
* Lines 63-68, creates the council area (category) rows for all financial years with the council area name the category split.
* Lines 70-75, HSCP category with the HSCP names the category split, for all financial years.
* Lines 77-83, Age/Sex category with a combination of the age group and sex as the category split, for all financial years.
* Lines 85-91, All ages/Sex category with age groups combined to ‘All ages’ and sex defined as the category split for all financial years.
* Lines 93-98, Age/All sex category with the patient sex combined to ‘Both’ to have Age group and Both as the category split, for all financial years.
* Lines 100-105, All ages/Both category with both age groups combined to ‘All ages’ and all genders combined to ‘Both’ to produce All ages/Both as the category split for all financial years.
* Lines 107-112, SIMD quintile as the category, with SIMD 1-5 as the category split for all financial years.
* Lines 114-119, SIMD top 15% as the category with the top 15% most deprived and other 85% as the category split for all financial years.
* Lines 121-126, Urban/rural 6 fold as the category and the 6 options being the category split, for all financial years.
* Lines 128-133, Urban/rural 2 fold as the category and the urban/rural option being the category split, for all financial years.
* Lines 135-144, includes the old length of stay calculation for comparison, with comparison as the category and ‘old’ marked as the category split for all financial years.
* Lines 147-238, this section of the code writes the data just produced into the template excel outputs.
* Line 151 reads in the figure template saved in the reference files folder.
* Lines 156-159, reads the data produced above into the ‘data’ tab, starting from column 2 as the first column is an index variable produced by combining financial year, category and category split. All figures should now be populated apart from figure 2 (map) which must be inserted using additional code.
* Lines 163-175, this section looks at the ‘calculation’ tab, defining whether or not the publication is either provisional or an update, which in turn completes the ‘Notes’ tab.
* Lines 180-186, this section inserts the picture produced from script 03, in this case ‘Figure 2’ which is the map, defining the width and height of the image and where it should be placed on the page. This has to be included as the map has to be created separately, all other figures should automatically populate once the data tab is filled in lines 150-153.
* Lines 190-196, adds a hyperlink to the note produced on row 19 column 3 of the ‘Notes’ tab and then hides the ‘Data’ and ‘Calculation’ tabs.
* Lines 200-202, saves the figures excel document in the output folder, adding the publication date to the file name.
* Line 206 reads in the qom template saved in the reference files folder.
* Lines 211-214, reads the data produced above into the ‘data’ tab, starting from column 2 as the first column is an index variable produced by combining financial year, category and category split. All data tables and charts should now be populated.
* Lines 216-230, this section looks at the ‘calculation’ tab, defining whether or not the publication is either provisional or an update, which in turn completes the ‘Notes’ tab.
* Lines 234-240, creates a hyperlink on row 16, column 3 of the ‘Notes’ tab and hides the ‘Data’ and ‘Calculation’ tabs.
* Lines 244-246, saves the qom excel file in the output folder, adding the publication date to the file name.
* The colour scheme for the 2 excel files ‘figures-template.xlsx’ and ‘qom-template.xlsx’ need to be fixed manually. Even if codes are correct in R, when the outputs get transferred from R to Excel something goes lost and Excel doesn’t retain the colour scheme specified in R.

05\_create-open-data

* Lines 18-25, reads in setup environment script and reads in basefile.
* Lines 28-68, computes Scotland total for health board file. Creates Scotland health board code and includes all years then aggregates the data (32-38). Then, add these rows on to the basefile, renaming variables for matching where required (40-46). Include provisional marker for latest financial year if required (48-53). Then re-order and rename variables (56-68).
* Lines 71-97, computes the HSCP file. Summarises basefile and then mark the latest FY data with the provisional marker where required (81-83). Re-order and rename variables (85-97).
* Lines 100-126, computes the Council Area file. Summarises basefile and then mark the latest FY data with the provisional marker where required (110-112). Re-order and rename variables (115-126).
* Lines 129-167, computes the Age/Sex file. Summarises basefile, mark the code for Scotland and then mark the latest FY data with the provisional marker where required (142-144). Modify ages and also missing sex to make look tidier in excel file (156-151). Re-order and rename variables (154-167).
* Lines 170-201, computes the deprivation file. Summarises basefile, mark the code for Scotland and then mark the latest FY data with the provisional marker where required (182-184). Tidy up deprivation quintiles for excel file to just show numbers 1-5 (186). Re-order and rename variables (189-201).
* Lines 203-231, computes the urban/rural file. Summarises basefile, mark the country code for Scotland and then mark the latest FY data with the provisional marker where required (214-216). Re-order and rename variables (219-231).
* Lines 234-260, creates open data folder and save each of the out files as a csv document into the folder using the required file name including the publication date.

06\_knit-markdown

* Details of this script are taken from two subset scripts, summary.Rmd and report.Rmd, which format the figures and data created in the previous scripts into a template summary and report document.

Summary.Rmd

* Lines 2-5 outlines the reference document to which the markdown script is being linked to. Line 5 removed the Alt Text that appears underneath each graph.
* Lines 10-18 reads in the setup environment and function scripts required as well as the basefile.
* Line 20-21 defines the name for the basefile for the most recent financial year to be inserted into the text and bullet points to be included in the summary file.
* Lines 23-32 creates a variable name for information broken by age and sex that will be used in the main points section.
* Lines 35-53 creates the alt text for the bar chart included in the main points section, using variables defined in previous scripts that ensure the correct definitions and numbers are inserted.
* Lines 55-58 creates a variable for each of the three contacts for this publication to be inserted further down the summary file.
* Lines 62-77 defines the headings for the summary file, including some code as to whether or not it is a provisional publication as well as the publication date.
* Lines 79-83 inserts the text for the ‘About this release’ section, including the necessary dates derived from the setup environment script.
* Lines 84-107 inserts the main points again with r code included to define specific figures and financial years previously defined. This is done for the first bullet points in lines 87-89. Lines 91-97 insert the bar chart and create the heading, there is also some script to include a note if this is a provisional release, if it is the October release however this note will not be included. Line 99-107 inserts the two final bullet point for this section using the same process as the first three.
* Lines 110-118 complete the Background information, inserting the appropriate hyperlinks.
* Lines 119-132 inserts the contact details for the publication created in lines 55-57 and in a specific format alongside each other and using the same font, colouring and spacing.
* Lines 134-146 inserts the ‘Further Information’ section again including publication links where necessary, hyperlinks and inserts the date of the next publication.

Report.Rmd

27/11/2023 PHS is working on new templates, there the templates saved in *‘…end-of-life-pub/markdown’* must be changed for the 2023/24 annual publication.

* Lines 2-5 outlines the reference document to which the markdown script is being linked to.
* Lines 9-16 reads in the setup environment and functions scripts required as well as the basefile, in addition to removing the package ‘tidylog’.
* Line 31 defines the name for the basefile for the most recent financial year to be used throughout this script which will insert required data into the report.
* Lines 34-36 defines names for a summarised dataset of the basefile by health board, as well as the creating variables for the maximum and minimum health board based on qom calculated in previous scripts.
* Lines 38-41 defines a name for a summarised dataset of the basefile by HSCP, as well as creating variables for the maximum and minimum HSCP based on qom calculated in previous scripts.
* Lines 43-51 looks at the combination of age and sex. Firstly, it defines a name for a summarised dataset of the basefile by sex and also creating variables for the youngest and oldest age groups where the variable ‘sex’ is missing from the summarised data source. Maximum and minimum variables are then created for each of the oldest and youngest age groups based on qom calculated in previous scripts.
* Lines 53-56 defines a name for a summarised dataset of the basefile by SIMD, as well as creating variables for the maximum and minimum simd based on qom calculated in previous scripts.
* Lines 58-66 defines a name for a summarised dataset of the basefile by urban/rural classification, as well as creating variables for the maximum and minimum urban/rural classification based on qom calculated in previous scripts.
* Lines 68-74 creates a provisional footnote for when this is required, inserting the latest financial year and future publication date where necessary.
* Lines 76-94 inserts the alt text for figure 1, including financial year values and trend information defined above and figures calculated in previous scripts.
* Lines 98-121 creates the completeness table for SMR01 by reading in the rds file saved in the extracts folder, broken down by health board, including a Scotland total and flagging if any values have a low completeness percentage, in this case if the value is less than 80%.
* Lines 123-148 creates the text for the table to be used in the metadata as well as the table to be used in the appendix, using the numbers and financial year definitions calculated above, with some additional definitions if the low percentage equals 0.
* Lines 151-201 creates the completeness table for SMR01 GLS that will be introduced for next year annual publication (2023/24) by reading in the rds file saved in the extracts folder, broken down by health board, including a Scotland total and flagging if any values have a low completeness percentage, in this case if the value is less than 80%.
* Lines 185-198 creates the text for the table to be used in the metadata as well as the table to be used in the appendix, using the numbers and financial year definitions calculated above, with some additional definitions if the low percentage equals 0.
* Lines 204-236 creates the completeness table for SMR04 that will be introduced for next year annual publication (2023/24) by reading in the rds file saved in the extracts folder, broken down by health board, including a Scotland total and flagging if any values have a low completeness percentage, in this case if the value is less than 80%.
* Lines 238-254creates the text for the table to be used in the metadata as well as the table to be used in the appendix, using the numbers and financial year definitions calculated above, with some additional definitions if the low percentage equals 0.
* Lines 259-277 inserts the publication information.
* Lines 284-294 inserts the information about Official Statistics publication, including links where necessary.
* Lines 324-326 inserts the text to be included on page 5 of the publication, including some script that defines the correct text to be included if it is a provisional release or a revised release.
* Lines 329-344 inserts the main points using the financial year, numbers and percentages that have been defined above.
* Lines 346-349 includes the beginning of the results and commentary section, inserting the first two paragraphs using financial years, numbers and percentages defined above within the text.
* Lines 350-364 inserts figure 1 as well as the title, including alt text and also a provisional footnote if required.
* Lines 399-375 inserts the text to introduce figure 2, using percentage figures and health board information defined in previous script where necessary.
* Lines 376-393 inserts figure 2 which is the map of Scotland split by NHS board of residence, with the necessary title as well as notes.
* Lines 395-397 introduces table 1 and inserts the title.
* Lines 397-417 creates table 1, summarising data from the basefile by health board, adding on a row for Scotland and selecting the measures to be displayed, which in this case are health board, number of deaths, qom (% time spent at home/in the community) and average number of days spent at home/in the community. Header labels are also created.
* Lines 421 -426 inserts the notes to summarise table 1.
* Lines 430-431 inserts the text which introduces table 2 and the title for the table.
* Lines 433-457 creates table 2, which summarises data from the basefile by HSCP, adding an additional row for Scotland and selecting the measures to be displayed, which are HSCP, number of deaths, qom (% time spent at home/in the community) and average number of days spent at home/in the community. Header labels are also created.
* Lines 461-470 are the notes which summarise table 2.
* Line 473 is a small piece of text linking to figure A1.2 from the appendix which is another relevant figure in addition to the tables above.
* Lines 477-490 inserts the text which introduces figure 3, the heading for the figure and then the bar chart itself, with any necessary notes below if a provisional publication.
* Lines 492-510 introduces figure 4 with some text, inserts the heading and the bar chart, as well as any notes below the chart. Finally, there is some additional text to summarise the chart.
* Lines 512-530 introduces figure 5 with some text, the header and the bar chart and any necessary notes below the chart. There is then some text below which summarises the chart.
* Lines 533-561 inserts the glossary.
* Lines 564-580 inserts the contact information for the analysts in the end of life team.
* Lines 582-602 inserts the ‘Further Information’ and ‘Rate this publication’ sections below the contact information.
* Lines 604-618 creates the heading for figure A1.1 and inserts the series of line charts for each health board and Scotland, as well as any relevant notes.
* Lines 620-630 goes through the same process for figure A1.2, which produces a line chart for each HSCP and additional notes.
* Lines 632-643 completes the same process again for figure A1.3, which produces a line chart for each simd quintile and additional notes.
* Lines 645-656 inserts figure A1.4 using the same process, which produces a line chart for each urban/rural classification and additional notes.
* Lines 658-676 inserts the background information text for appendix 2.
* Lines 678-684 inserts the text which introduces table 3.1, 3.2 and 3.3 that will be included for next year publication 2023/24 on data completeness, using the notes and percentages defined in creating the completeness table above, as well as any required links.
* Lines 686-706 creates the title for table 3.1 SMR01 and inserts the table, which was created in an above script, including header labels.
* Lines 709-730 creates the title for table 3.2 SMR01 GLS and inserts the table, which was created in an above script, including header labels. IMPORTANT lines 723-726 analyst must contact Data Quality Team to check with quarter is available and which one is not because the code will stop running when you run the code at the end of August (Susan Law’s Team).
* Lines 738-752 creates the title for table 3.2 SMR04 and inserts the table, which was created in an above script, including header labels. IMPORTANT lines 746-749 analyst must contact Data Quality Team to check with quarter is available and which one is not because the code will stop running.
* Lines 755-757 inserts the text for appendix 3, which is the methodology section.
* Lines 761-771 creates the calculation of measure table, using three columns, listing the x labels which is the calculation measure, the y label splits the x and z columns with an equals sign, with the z column containing the text of how to complete the calculation.
* Lines 773-795 inserts the text to introduce figure A3.1, as well as including figure A3.1 with the alt text to be used and any necessary notes.
* Lines 800-803 inserts the Revision of methodology part.
* Lines 805-929 inserts the metadata table created in the above script, split into two columns named metadata indicator and description.
* Lines 931-945 completes appendix 5, which is the early access details section.
* Lines 947-953 completes appendix 6, which is the PHS section, again this is mainly text orientated and is the final section in the publication.

**Open data files for the Scottish Government.**

This needs to be updated after the publication has been released. Normally SG people get in touch to have this file via PEoLC Inbox.

See R codes from \\stats\irf\18-End-of-Life\Publication\SG open data to produce the output.

All previous communication can be found in the ‘SG open data’ subfolder within the PEoLC Inbox, as well as the instructions on how to upload this file.