# Scottish Household Survey
# Statistical Disclosure Control of 2019 Microdata
# - UK Data Archive –

# Version Control

| Version Number | Drafted by | Published / Draft |
|---|---|---|
| V0.1 | Darren Knox, SHS Senior Assistant Statistician, Communities Analytical Division | Draft |
| V1.0 | Darren Knox, SHS Senior Assistant Statistician, Communities Analytical Division | Published |
| | | |

# Document Approval

| Approved by | Role | Date |
|---|---|---|
| Alison Stout | SHS Project Manager, Communities Analytical Division | 21/01/2021 |
| | | |

# 1. Background

**1.1** The Scottish Household Survey (SHS) is a continuous survey based on a sample of the general population in private residences in Scotland. The survey started in 1999 and up to 2011 followed a fairly consistent survey design. From 2012 onwards, the survey was substantially redesigned to include elements of the Scottish House Condition Survey (SHCS) including the follow-up Physical Survey component. The survey is run through a consortium led by Ipsos MORI.

**1.2** The SHS is designed to provide reliable and up-to-date information on the composition, characteristics, attitudes and behaviour of private households and individuals, both nationally and at a sub-national level, and to examine the physical condition of Scotland's homes. It covers a wide range of topics to allow links to be made between different policy areas.

**1.3** The specific aims of the survey are to:

**1.3.1** provide household and individual information not currently available in Scotland, particularly to support the work of Communities, Local Government and Transport policy areas and the work of the Scottish Parliament;

**1.3.2** permit disaggregation of such information both geographically and in terms of population sub-groups (such as families with children or the elderly);

**1.3.3** allow the relationships between social variables within households to be examined. This will support cross-departmental and inter-departmental policies such as those on social justice and welfare-to-work;

**1.3.4** allow early detection of national trends;

**1.3.5** allow detailed follow-up surveys of sub-samples from the main survey sample, if required.

**1.4** In order to allow the survey to achieve best value from the data that is collected, a version of the data is made available on the UK Data Archive to allow researchers to carry out their own analysis. The data is released under End User Licence, and therefore the Scottish Government has taken action to protect the data from disclosure.

# 2. Introduction

**2.1** This document sets out the main ways in which the SHS data has been protected for risk of disclosure, or what is known as statistical disclosure control (SDC).

**2.2** This includes action taken on the 2019 fieldwork year's data, including:

**2.2.1** SHS social survey data; and

**2.2.2** Travel Diary data – journey and stage files including distance to school and distance to work datasets.

**2.3** The 2019 physical survey data (known as the Scottish House Condition Survey module) is out of scope of this document but will be made available under End User Licence in the future.

**2.4** The Government Statistical Service (GSS) Disclosure Control Guidance for Microdata Produced from Social Surveys (October 2014)[1] cites principle 5 of the Code of Practice for Official Statistics[2] for guidance on the publication of microdata under [End User Licence] EUL, requiring that: "official statistics do not reveal the identity of an individual or organisation, or any private information relating to them, taking into account other relevant sources of information."

**2.5** "Information relating to an individual should be considered by a producer of statistics to be 'private' if it was provided with the expectation that the information would be kept out of the public domain. Survey pledges provide respondents with assurances that the information they provide will remain confidential."

**2.6** Furthermore, the guidance states that "arrangements for confidentiality protection are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics".

**2.7** Private information is defined in the National Statistician's Guidance; Confidentiality of Official Statistics as being information that:

**2.7.1** relates to an identifiable legal or natural person, and

**2.7.2** is not in the public domain or common knowledge, and

**2.7.3** if disclosed would cause them damage, harm or distress

**2.8** Previous SHS sweeps up to the two year fieldwork years 2007/2008 received limited disclosure control. The focus of the disclosure control was around a small amount of top-coding and banding of certain identifying variables.

**2.9** This document sets out the main areas of protection which have been applied to the 2019 SHS social survey data which is deposited on the UK Data Archive[3]

[1] Policy for social survey microdata
[2] It should be noted that this guidance does not refer to the most recent version of the Code of Practice for Statistics. In the most recent version this principle is covered by the Data governance principle under the Trustworthiness pillar.
[3] http://data-archive.ac.uk/

# 3. Statistical Disclosure Control Methods

**3.1** For the 2012 SHS main social data (including the travel diary module) a three step process was adopted to reduce the risk of disclosive information being released. The **first step** identified and processed key variables to reduce their risk to disclose private information about respondents. The **second step** was to conduct an intruder test on the amended dataset. This produced a set of recommendations which were implemented in **step three** to produce the final data set described in this report. The resulting disclosure control has been applied to the 2019 SHS main social data (including the travel diary module). All new variables appearing in 2019 have been assessed separately.

**3.2** Three methods were used to disclosure control the main data set: re-coding, top-coding and variable removal.

**3.2.1** **Re-coding** is used on categorical variables to collapse response categories together into larger groups to hide small number of respondents in some of the rarer groups (e.g. grouping together rarer ethnicities into broad categories).

**3.2.2** **Top (or bottom) coding** is used on numerical variables where values of a variable above or below a specified level are grouped together and labelled as such. For example, age may be top-coded at 86 and reported as '86+'.

**3.2.3** **Variable removal** (or suppression) is used to remove sensitive variables (e.g. sexual orientation). In addition, variables used to calculate other summary variables (e.g. household relationships are used to derive household type) can be disclosive if not recoded or suppressed. These can help in identification of respondents in combination with other variables.

**3.3** The following sections of this report outline the set of SDC approaches applied to variables in the 2019 SHS Main Dataset (including the travel diary module), using each of these three methods.

# 4. Suppressed Variables

**4.1** A number of variables have been removed (suppressed) from the EUL datasets.

| Topic | Variable | Notes |
|---|---|---|
| **Admin** | Date of interviews, interview stream, agreement to carry out physical survey (streams 1 to 4 only), random adult interview<br><br>QUARTER<br>FYEAR<br>H_DATE, DAY, MONTH AND YEAR<br>R_DATE, R_DAY, R_MONTH AND R_YEAR<br>STREAM<br>PS1<br>HHPERM<br>RANDPER<br>P2SAMEDAY | Interviews are batched up by area to minimise interviewer travel costs. The interview date could be used to identify the respondent's local area. |
| **Geography** | Scottish Index of Multiple Deprivation (SIMD) 2020 deciles, 15% most deprived vs. rest of Scotland and older versions of SIMD<br><br>MD12DEC<br>MD12QUIN<br>MD12PC15<br>MD16DEC<br>MD16QUIN<br>MD16PC15<br>MD20DEC<br>MD20PC15<br><br>Six-fold urban rural classification 2016<br><br>SHS_6CLA1314 | SIMD 2020 quintiles and two-fold urban rural classification 2016 (re-coded) are provided (see section 5, re-coded variables). |
| **Household Matrix** | Relationships between household members<br><br>HA3_1_2 TO HA3_10_9<br><br>OTH1REL TO OTH3REL<br><br>NUMBHH HHSIZE TOTPEEPS | The inclusion of these household composition variables can negate the purpose of summary variables and provide a disclosive risk of identifying large households or other uncommon household types from the data set. |

| Topic | Variable | Notes |
|---|---|---|
| | | Any household of more than 6 people was removed from the dataset as they are intrinsically disclosive. |
| **School Child** | School of random school-age child<br><br>SCHOOL | School of random school-age child is highly disclosive for some households in certain local authorities. |
| **Health and Disability** | Health conditions of household members<br><br>HF2A_P1_01 TO HF2A_P10_20 | Health conditions of household members considered sensitive and could identify rare households. |
| **Economic Activity and Education** | Economic activity, Number of jobs/hours worked, Employment, National Statistics Socio-Economic Status<br><br>ILO1_1 TO ILO1_6,<br>ILO2 TO ILO4,<br>RAILO1_1 TO RAILO1_6,<br>RAILO2 TO RAILO4,<br>HG18, HIHWRK2, HG22 TO HG28,<br>HH1 TO HH55,<br>HH1_DKREF TO HH54_DKREF,<br>NSSEC<br>EDQUAL01 TO EDQUAL13,<br>OA1INEMP TO OA3INEMP,<br>OA1EARN2 TO OA3EARN6,<br>EARNSUM_OA1 TO EARNSUM_OA3,<br>EARNSUM,<br>EARNINC,<br>EARNINC_OA1 TO EARNINC_OA3 | Variables that are used to generate summary variables which could be used to disclose information about an individual have been removed.<br><br>Dataset contains a number of summary variables in place (e.g. NSSEC2). |
| **Household finances** | Component amounts of net annual income from benefits and miscellaneous sources, summary variables for different components of income (describing level of imputation, if any) and amount<br><br>Income from benefits<br><br>• All variables beginning HH56, HH57 or HH60 **that relate to adults other than the highest income householder and their spouse** | Many income variables are collected during interview in order to derive imputed values based on the “total net annual” income.<br><br>Main social data includes the main income variables available (total net annual income, income from earning, benefits, miscellaneous sources).<br><br>Many summary variables have been included (e.g. BENINC_HIHSP; total net annual income from benefits for the highest income householder and their spouse) |

| Topic | Variable | Notes |
|---|---|---|
| | • All variables beginning HH61 or HH62<br>• All variables beginning BENSUM **except** BENSUM_HIHSP<br>• All variables beginning BENINC **except** BENINC_HIHSP<br><br>Income from miscellaneous sources<br><br>• All variables beginning HH67 to HH71<br>• All variables beginning MSCSUM **except** MSCSUM_HIHSP<br>• All variables beginning MSCINC **except** MSCINC_HIHSP<br><br>Derived variables on earnings from main and other jobs for the HiH and their spouse<br><br>HINCMINC, SINCMINC, HINCOINC, SINCOINC, HINCMSUM, SINCMSUM, HINCOSUM, SINCOSUM<br><br>Other derived variables on household finances<br><br>WKNETINC, WKNETINC_BROAD, BANDINC, BANDINC_BROAD, INCBAND, INCBAND_BROAD, ANNETINC_BROAD, INCSUM_BROAD, TOTHINC_BROAD<br><br>Council tax band and total and after deductions council tax payments<br><br>COUNCILTAXBAND<br>CT2A<br>CT2B | |

| Topic | Variable | Notes |
|---|---|---|
| **Housing** | Mortgage/rent payments<br><br>HJ2 to HJ6 | These variables have been excluded from the dataset. However, the new derived variables on housing costs (e.g. mortgage_amt and rent_amt) are provided. |
| **Transport and Travel** | Name of school/college/university, location of workplace<br><br>RD1 RD2 | Name of school/university and location of the place where you work is highly disclosive for some households in certain local authorities. |
| **Composition / Characteristics** | Sexual orientation<br><br>RAND_ID | Deemed sensitive for release of EUL datasets |
| | Reason for discrimination or harassment being 'your sexual orientation' or 'your trans status, including non-binary identities'.<br><br>DISCRIM_09, DISCRIM_10, DISCRIM_15, HARASS_09, HARASS_10, HARASS_15 | High level of association between reason for discrimination and/or harassment being your sexual orientation and respondent identifying as lesbian/gay.<br><br>Small number of respondents reporting the reason for discrimination and/or harassment being your trans status, including non-binary identities.<br><br>Therefore these have been combined with the 'other' reason discrimination and harassment variables to create new variables DISCRIM_091015 and HARASS_091015 |
| | Gender<br><br>HIHGENDER<br>RANDGENDER<br>SPGENDER<br>KIDGENDER<br>OTH1GENDER<br>OTH2GENDER<br>OTH3GENDER<br>GENDER<br>HA6_1 TO HA6_10 | One tenth of one per cent (0.1 per cent) of adults described their gender as 'in another way' or did not provide an answer. As re-coding is not appropriate for the gender variables, any households where a respondent has described their gender as 'in another way' or did not provide an answer have been removed. |
| **Number of Children** | Total number of children and total number of eligible children in the household<br><br>TOTKIDS<br>NUMKIDS | Households containing 5 or more children have been removed from the data because households with large numbers of children present are easily identifiable. This affects approximately 20 households. |
| **Number of Adults** | Total number of adults and eligible adults in the household<br><br>TOTADS<br>NUMADS | Households containing 6 or more adults have been removed from the data because households with large numbers of adults present are easily identifiable. This affects approximately 10 households. |

| Topic | Variable | Notes |
|---|---|---|
| **Childcare and childcare costs** | Childcare costs and childcare involving overnight stays<br><br>ELC2A<br>ELC2B<br>ELC3A<br>ELC3B<br>ELC11<br>ELC11A<br>ELC12<br>ELC12A<br>ELC17_KIDOTH1 TO<br>ELC17_KIDOTH5<br>ELC17A_KIDOTH1 TO<br>ELC17A_KIDOTH5<br>ELC18_KIDOTH1 TO<br>ELC18_KIDOTH5<br>ELC18A_KIDOTH1 TO<br>ELC18A_KIDOTH5<br>ELC20<br>ELC20A<br>ELC20B | A small number of households reported childcare arrangements involving overnight stays. The variables allowing these households to be identified have been excluded.<br><br>Childcare costs are reported as either weekly or monthly. Therefore, there is no easy way of top coding childcare costs so these variables have been excluded. |
| **Redundant** | Council tax instalments<br><br>CT3 | After excluding the council tax band and council payment variables CT2A and CT2B which are disclosive (see housing topic), CT3 becomes redundant |
| | Age of respondents/spouse/random school child<br><br>HIHAGEBD<br>SPAGE<br>AGERANK<br>KIDAGE<br>AGERBAND<br>OTH1AGE TO OTH3AGE<br>AGE_7 TO AGE_10<br>PENAGE_7 TO PENAGE_10<br>HA6_7 TO HA6_10<br>HA7_7 TO HA7_10<br>HA10_7 TO HA10_10<br>HA13_7 TO HA13_10<br>HF1A_07 TO HF1A_10 | Single year of age provided for highest income householder and random adult.<br><br>Banded age is provided for all other persons.<br><br>Any household of more than 6 people was removed from the dataset as they are intrinsically disclosive. Therefore the variables AGE, PENAGE, HA6, HA7, HA10, HA13 for persons 7 to 10 become redundant. |
| | NUMCARS | A new variable NUMCARS_NEW is provided which details number of cars available to household (NUMCARS included number of vans as well) |

| Topic | Variable | Notes |
|---|---|---|
| | HA6_P1_OTHERWAY TO HA6_P10_OTHERWAY | See the notes on the gender variables in the composition and characteristics topic. |

# 5. Recoded Variables

**5.1** Recoded variables are usually categorical variables. In most cases this involves combining smaller categories to reduce the risk of disclosure of a household adult.

| Topic | Variable | Notes |
|---|---|---|
| **Admin** | Serial number, survey weights<br><br>UNIQID<br>LA_WT<br>IND_WT<br>KID_WT<br>TRAV_WT<br><br>LA_GRWT<br>IND_GRWT<br>KID_GRWT | The unique identifier (serial) variable is recoded to hide information about the order the interviews were carried out in, as this can be linked to local areas. This recoding is consistent with the travel diary module.<br><br>Survey weights are rounded to two decimal places. |
| **Composition / Characteristics** | Religion and Ethnicity<br><br>HIH_REL<br>RNDREL09<br>HIH_STAT<br>RANDSTAT<br>HIH_ETH<br>RANDETH | Religion and ethnicity are key variables that could lead to the identification of small groups of people.<br><br>Religion has been recoded to group all non-Christian religious groups together, due to the small numbers of respondents in these groups. Ethnicity was recoded to group small white ethnic groups together and non-white minority ethnic groups together..<br><br>Marital status of highest income householder and random adult are coded to bring together similar groups of marriage and civil partnership (e.g. married / civil partnership, divorced/legally dissolved civil partnership) |

| Topic | Variable | Notes |
|---|---|---|
| | Country of birth<br><br>RANDFROM | Due to small base sizes for UK countries (excluding Scotland) and the Republic of Ireland as the birth country of the respondent, the country of birth is a key variable which could lead to identification of respondents. The UK countries excluding Scotland have been grouped together and the Republic of Ireland and the Elsewhere category have also been grouped together. This approach is consistent with the pooled sample. |
| | AGE_1 to AGE_6 | Ages of individuals in the household have been banded by rough school age groups (0-4, 5-11, 12-15) and then five year age bands. |
| **Geographic** | Six-fold urban rural classification 2016 and Scottish Index of Multiple Deprivation (SIMD) 2020 quintiles<br><br>SHS_6CLA<br>MD20QUIN | SHS_6CLA (six-fold urban rural classification 2016) has been recoded to the two-fold (urban or rural) classification (SHS_2CLA) as it could disclose unique data zones of respondents.<br><br>A small number of data zones can be uniquely identified by combinations of council, SIMD 2020 quintile and two-fold urban rural classification 2016. In these cases the urban rural classification was changed. Generally this involved changing the classification from rural to urban. If this resulted in an implausible combination then the SIMD quintile was changed. |
| **Housing** | Accommodation type<br><br>HB1<br><br>Rent payments<br><br>RH4 and RH8 | Some rare types of accommodation have been coded together.<br><br>Questions on the number of people in the household contributing to rent payments and on what other costs rent payments included have been recoded. |
| **Transport and Travel** | Main mode of travel (in the travel diary dataset)<br><br>MAINMODE | Smaller categories of mode of travel have been combined |
| | Two-fold urban/rural classification for journey/stage start/end points (in the travel diary dataset)<br><br>JOURNEYSTART_RURURB<br>JOURNEYEND_RURURB<br>STAGESTART_RURURB | Unique combinations of LA/SIMD/Urban rural have been re-coded for consistency with main social data. |

| Topic | Variable | Notes |
|---|---|---|
| | STAGEEND_RURURB | |
| **Economic Activity / Employment** | Socio-economic variables<br><br>HIHSOC<br>HIHSIC | Banded together into broad analytical classes. |
| **Young People** | Local Authority of the school the children go to<br><br>SCLCOUN<br>Laschool (travel diary) | Recoded to group together where the local authority of a school is different to the household local authority, as this could lead to the identification of rare households. |
| **Childcare and childcare costs** | Hours of unfunded/funded childcare per week<br><br>ELC2<br>ELC3<br>ELC4<br>ELC8 | Recoded to into the following groups to reduce the risk of identifying respondents with unique childcare hours.<br><br>None/Less than 1 hour<br>1-10 hours<br>11-20 hours<br>21-30 hours<br>Over 30 hours |

## 6. Top (or Bottom) Coding Variables

**6.1** Many numerical variables may identify households/adults due to values that are unusual or very low/high (e.g. very young/old adults).

| Topic | Variable | Notes |
|---|---|---|
| **Household Composition / Characteristics** | Age of highest income household and random adult<br><br>HIHAGE<br>RANDAGE | Top-coded at 86+.<br>HIHAGE also bottom coded at '18 and under' |
| | Year of arrival in UK<br><br>MIGBIR2_Y | Reference to very old years may identify people who arrived in the UK a very long time.<br>MIGBIR2_Y is top-coded at 30+ years. |
| **Housing** | Date first started renting/own property<br><br>T38YEAR<br>T39YEAR<br>RA1 | Reference to very old years may identify people that have lived in areas for a very long time.<br>T38YEAR, T39YEAR, and RA1 are top-coded at 30+ years. |
| | Number of bedrooms<br><br>HC4 | May lead to identification of very large, rare, households. Top-coded at 5+. |
| | Number of housing lists on | May lead to the identification of rare households.<br>Top-coded at 2+ |

| Topic | Variable | Notes |
|---|---|---|
| | RA15A | |
| | Purchase price/Amount borrowed and housing costs<br><br>OH9<br>OH14<br>OH15<br>OH16<br>MORTGAGE_AMT<br>RENT_AMT<br>SHARED_OWNERSHIP_AMT<br>HCOST_AMT | May lead to identification of very expensive households or households with very high costs. Top-coded at 95th percentile and bottom-coded at the 5th percentile. |
| **Transport and Travel** | Number of cars, amount spent on fuel, number of bicycles<br><br>NUMCARS_NEW<br>NUMVANS<br>HD19<br>HD2011 | May lead to identification of households with very large number of cars, offers limited utility. NUMCARS_NEW top-coded at 2+ and numvans top-coded at 1+.<br><br>Amount spent on fuel very high for some individuals. Top-coded at 95th percentile.<br><br>Number of bicycles top-coded at 3+. |
| **Household Finances** | Income from benefits, miscellaneous sources and earnings.<br><br>BENINC_HIHSP<br>MSCINC_HIHSP<br>EARNINC_HIHSP<br>ANNETINC | Very large incomes should be protected as it may lead to identification of households and the people within them.<br><br>All income variables are top-coded at the respective 95th percentile. |