Working Paper 2005/2

SDC-i

SOFTWARE MANUAL V2.0

Software for assessing the impact of statistical disclosure controls on end-user analyses

Paul Williamson

June 2005

Population Microdata Unit
Department of Geography
University of Liverpool

Contents

Introduction for first time users

SDC-i is a software suite aimed at helping to assess the impact of statistical disclosure control on end-user analyses. Figure 1 (p.4) illustrates the logic flow of the program suite. However, each main element can also be run as stand-alone module. For example, users with their own set of pre- and post-adjustment cell counts can use the *SDC_Direct_Impacts* module to measure the impacts of adjustment without having to run any of the other modules.

SDC_Direct_Impacts has been written to allow maximum flexibility over the location of program files, including all input data. For all other modules, references in this manual are given relative to *D:\Research\Disclosure Control\Disclosure_VB*.

Quick Start Guide

For most users, the main program of interest will be *SDC_Direct_Impacts*. To get the most out of this package it will be necessary to read pages 18-41 of manual. However, the basic functionality of the program can be mastered will less effort:

(1) Download zipped executable version

(2) Unzip package (includes executable code, default program parameters, example benchmark data and copy of user manual)

(3) Double click on program to run (to check program works on system) (run-time c. 2-4 mins)

(4) Examine files in folder *SDCi Input Counts* containing example pre- and post-perturbation counts; use as template for formatting own input data. Name each file using the convention <*table name*>**_v*n*.fmt** , where *n* = 0 if pre-perturbation of counts and *n*=1 for post-perturbation variant. (e.g. *UserTable_v0.fmt*)

(5) Read pages 21-24 of manual, explaining steps necessary for creation of table mappings.

(6) In the *Parameters* folder edit the file *SDC_Direct_Impacts_Count_input_tables* to list instead user supplied table(s) (see pages 35-36 (section 6) of user manual for details.)

(7) Run program; results of comparison will be placed in file *SDC_Direct_Impacts_results.txt*

(8) Change user parameters to request alternative summary measures as required and re-run program (see pages 24-35 of user manual for details.)

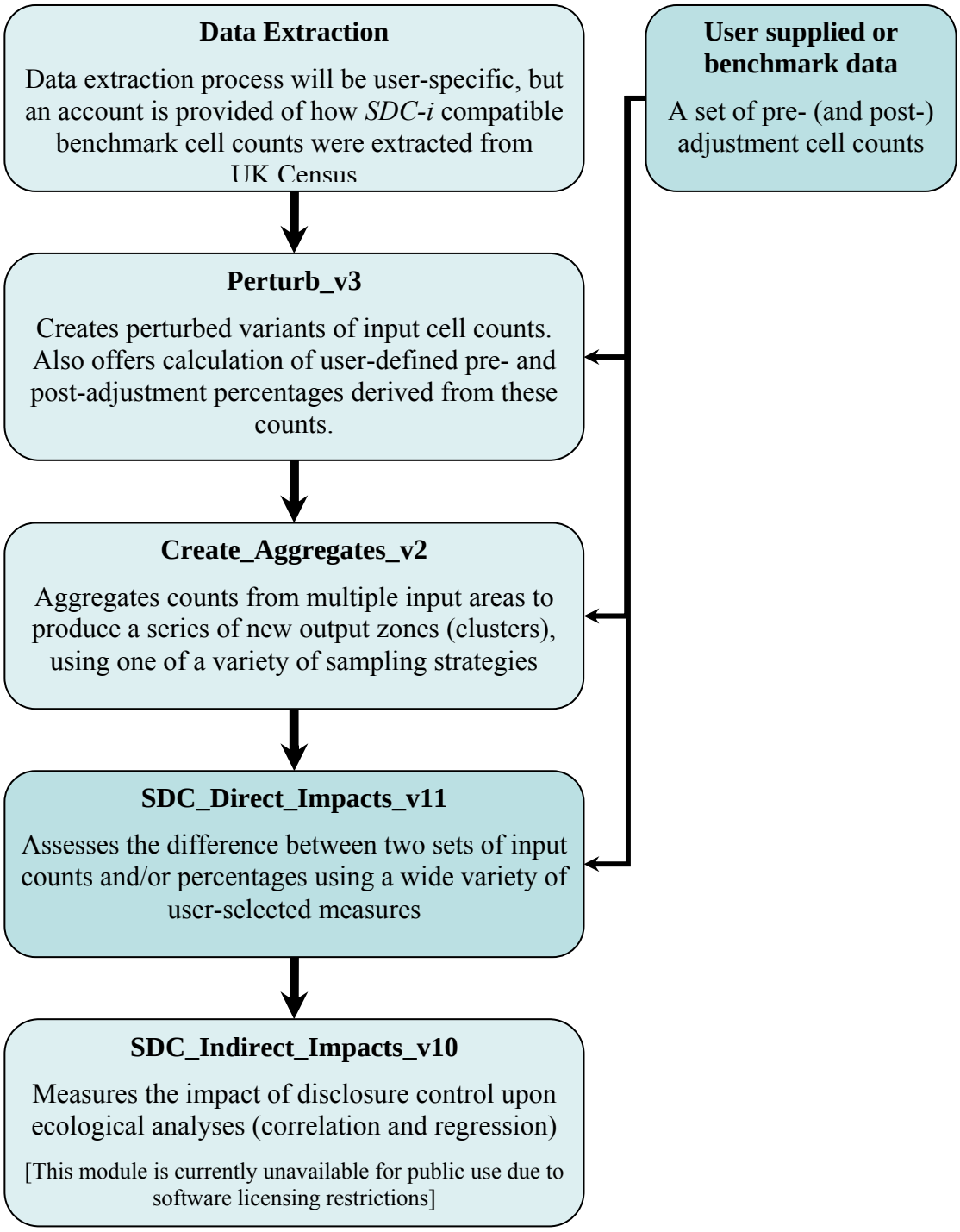

Figure 1 Linkage between SDC-i modules

Data extraction

Details of how to extract a set of pre-perturbation cell counts (including marginals), for a set of user-specified areas, will obviously be specific to the data source and data access route used. The following details the process used to extract a set of counts from the 1991 UK Census via SASPAC, a PC-based census data extraction tool. The precise details of this process will be of no interest to the general reader, and are included here for internal documentation purposes. However, the broad steps followed, indicated by the numbered headings and comments in bold below, offer some guidance as to the general approach required. The file *threshold.fmt*, mentioned in step (iv), is required as an input to *Create_Aggregates* and *Perturb*; more details may be found in the relevant sections of the user manual.

(i) Identify representative sample of EDs

The user-supplied sample of input areas should ideally be geographically representative. If it is intended to create aggregations of these areas, a spatially clustered sampling strategy is recommended.

The data used in the ED selection process is stored in the folder *Census_1991_ED_ pops* :

- *Selected_1991_EDs.xls* – file containing list of all non special/restricted/shipping EDS in England and Wales that meet intended min. population threshold of 100 persons/40 households; from which a shorter list, comprising 5% of EDs are chosen using population-weighted random selection
- *Selected_1991_EDs.csv* – file containing comma-separated list of chosen 'blocks' of 50 consecutive EDS, giving start and end ED of each block
- *Selected_1991_EDs.txt* – file containing above *.csv* list converted into SASPAC friendly format

Final result (*Census_1991_Selected_Wards.sav*): 107 selected blocks of 50 EDs (5.00% of the 2139 possible blocks), containing 5350 above pop. threshold EDs (5.00%); in turn containing a resident population of 2,522,173 (5.09%).

(ii) *Extract census table and strata data for these EDs*

If stratified sampling is required (i.e. the aggregation of input areas from similar social strata), then an additional set of data containing the relevant stratifying measure for each input area should be created.

SASPAC for PC cannot extract data for all England & Wales counties at the same time. Consequently, four command files are required per table, each processing 10 county system files. Command files saved in table-specific folders under *c:\saspac\command*. Raw output files, also stored in table-specific folders under *c:\saspac\interfac*. The four output **.fmt* files per table have been stitched into one using the *copy tablename_*.fmt tablename.fmt* command in DOS. Copies of these combined files have been placed in *SAS_Inputs*, renamed as *tablename_saspac.fmt*.

A further system limitation of SASPAC for PC means that attempting to exclude special, restricted and shipping EDs from output makes for very slow running. To save time, these exclusions were made only during the generation of the *threshold.fmt* files.

(iii) Drop special, restricted and shipping EDs; add marginal to S08

If required, input areas failing to meet specified threshold criteria should be dropped.

The program *Remove_invalid_EDs.f95* (in *SAS_Inputs/Program* folder) uses the list of valid EDs from *SAS_Inputs/threshold.fmt* (i.e. all EDs not special, restricted or shipping) to identify and strip out all invalid EDs from other *tablename_saspac.fmt files*, with results data for valid EDs only being written to *tablename_raw.fmt* files in same folder. N.B. At this stage, an ED's validity does not depend upon number of persons or households in ED meets 2001 Census threshold criteria; under 2001 Census pop. threshold EDs are stripped out by *Create_Aggregates*. This allows maximum flexibility for setting alternative threshold levels (although contiguous blocks of above threshold EDs have been designed to ensure 50 EDs remain once 2001 pop. thresholds have been applied).

For SAS Table 08 an alternative version of the program, *Correct_s08.f95*, is required, as not only do invalid EDs have to be stripped out of *s08_saspac.fmt*, but a total persons marginal has be calculated and added on the basis of the reported male and female totals.

(iv) Create Visual Basic friendly version of threshold.fmt

If input areas are to be dropped on the fly by *Perturb* or *Create_Aggregates*, a suitably formatted version of the datafile threshold.fmt will be required.

As extracted via SASPAC, *threshold.fmt*, unlike the other **_raw.fmt* files, does not store ED names on a separate line from the related count data. This is problematic for *Perturb*, *Create_Aggregates* and *SDC_Direct_Impacts*, as all are written in Visual Basic. In Visual Basic, any line who's first non-blank character is a non-numeric character is treated as a single string. E.g. 01AAAA01 20 30 is treated as one block of text, rather than an ED identifier followed by two counts. To get round this, the ED name has be put in quote marks, and the text and counts separated by commas. This has been accomplished using *Excel* (to create a *csv* file, with quote marks added to start and end of ED name [=”””””&A1&”””””]) and *Notepad* (to strip out excess quote marks apparent only after saving file and opening in other than Excel). The VB friendly version of *threshold.fmt* is called *threshold.fmt*, the original version being renamed *threshold_f95.fmt*. Both are located in the *SAS_Inputs* folder.

Perturb_v3

[In folder *Perturb_v3*]

Perturb creates perturbed variants of a set of initial table counts for above threshold areas. Each set of perturbed counts is output as a table variant. At present five table variants are produced:

(i) Rounding to base 5: stochastic rounding of all counts to base 5, with p=4/5 of being rounded to nearest multiple of 5 for remainders of 1 and 4; p=3/5 for counts of 2 and 3.

(ii) Rounding to base 3: stochastic rounding of all counts to base 3, with p=2/3 of being rounded to nearest 3.

(iii) Rounding of small numbers to base 3: stochastic rounding of all counts < 3 to base 3, with p=2/3 of being rounded to nearest 3.

(iv) Barnardisation: +/-1 on all cells, with +1 and –1 each having p=0.1

(v) Barnardisation: +/-1 on all cells, with +1 and –1 each have p=0.02

If the initial table counts are inconsistent (marginal <> sum of internal cells), *Perturb* corrects this inconsistency prior to perturbation.

Program limits

No. of input tables: any
No. of input areas: any
No. of table rows: 40
No. of table columns: 40
Area names: max. of 8 characters
Perturbation methods: limited to those hard-coded in the software (although code could be amended to incorporate others)

Program Inputs

1) *Perturbation input tables.txt*
[Stored in folder Perturb_v3]

Lists input tables, one per line, starting with table name in quotes (less file extension), followed after a comma by a flag indicating whether or not the table marginals as initially supplied are consistent with internal table counts (0=consistent, 1=inconsistent). Before perturbation a table flagged as inconsistent will have its initial marginal counts adjusted to fit its initial internal counts

Inconsistent table marginals can arise when multiple tables are concatenated. For example, in SAS Table 6, the count of Total Persons always sum to equal the stated counts of males and females (SAS Table 6[a]), but does not always sum to equal the stated counts by age group.

SAS Table 06(a) Ethnic group of Residents by Age and by Sex

Enumeration District: BYFA01

Sex and Age	Total Persons	**Ethnic group**										Persons born in Ireland
		White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups: Asian	Other groups: Other	
Total Persons	115	94	4	0	0	3	0	0	12	0	2	7
Males	54	45	1	0	0	0	0	0	6	0	2	1
Females	61	49	3	0	0	3	0	0	6	0	0	6

SAS Table 06(b) Ethnic group of Residents by Age and by Sex

Enumeration District: BYFA01

Sex and Age	Total Persons	Ethnic group										Persons born in Ireland
		White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups		
										Asian	Other	
Total Persons	115	94	4	0	0	3	0	0	12	0	2	7
0-4	6	5	0	0	0	0	0	0	1	0	0	0
5-15	5	3	0	0	0	0	0	0	2	0	0	0
16-29	52	44	1	0	0	0	0	0	5	0	2	5
30<pa	42	36	2	0	0	3	0	0	1	0	0	0
Pa and over	9	5	0	0	0	1	0	0	3	0	0	2

For these two tables, appropriate entries in the file *Perturbation input tables.txt* would be:

```
"s06a", 0
"s06b", 1
```

2) *<tablename>.map*
[stored in folder *Perturb_v3*]

For each table to be perturbed a file is required that maps internal table cells on to table marginals. This file, *<tablename.map* : file describing table layout, including (i) number of rows and columns; (ii) row counts which sum to give row marginal(s) (one mapping per marginal) [if any]; (iii) column counts which sum to give column marginal(s) (one mapping per marginal) [if any]. For full details, see description of mapping process under description of inputs to the main *SDC_Direct_Impacts* program.

3) *threshold.fmt*
[stored in folder *SAS_inputs*]

For each input area, the file *threshold.fmt* records the following information: Name of area (in quotes), number of persons, number of households. One line is used per area. To comply with input format requirements, the three items must be comma-separated, with the area name in quotes.

e.g.

```
"04BXFA01",458,136
"04BXFA02",280,83
```

The file *threshold.fmt* is used as the definitive master list of input area names; other input files presenting area-level data in a different order will cause execution errors that will be flagged in program output. The information on persons and households is used by *Create_Aggregates* and *SDC_Direct_Impacts* to identify those input areas that fall below minimum threshold levels (of households and/or persons).

The data from *threshold.fmt* can be extracted using SASPAC, but post extraction Excel or similar must be used to ensure area names are in quote marks, otherwise *Perturb* will treat the whole line as the area name.

4) *<tablename>_raw.fmt*
[stored in folder *SAS_inputs*]

For each table to be perturbed, a set of initial table counts is required. These counts can be laid out in vector or matrix format according to user preference. In the file, the name for each area is followed on subsequent lines by the relevant table counts, space separated.

e.g.

```
04BXFA01
 458  0  453  5  134  2
 50  0  50  0  13  0
04BXFA02
 280  0  272  8  81  2
 24  0  24  0  8  0
```

The table layout, including number of rows, number of columns and mapping of internal onto marginal table cells, is supplied as an input to the program via the *<tablename>.map* files described above. Data should be presented for every area listed in *threshold.fmt*, in precisely the same order.

If used, *Convert_SAS* automatically converts SAS tables extracted in vector format from CASWEB into matrix format (i.e. row x column).

Program Outputs

1) *Table mapping check.txt*
[stored in folder *Perturb_v3*]

The correct mapping of internal table counts to table marginals is vital to help ensure that, where appropriate, table marginals are correctly recalculated to reflect perturbations to internal cell counts. To help check the table mappings supplied as inputs by the user in the *<tablename>.map* files, *Table mapping check.txt* reports the number of internal cells that each table cell count is derived from.

e.g.

```
Table Mapping for table63
   28   21    7    7    7    7
    4    3    1    1    1    1
    4    3    1    1    1    1
    4    3    1    1    1    1
    4    3    1    1    1    1
    4    3    1    1    1    1
    4    3    1    1    1    1
    4    3    1    1    1    1
```

In the above example, the top row and first two columns are table marginals. The grand table total (top left-hand corner) is derived, therefore, from the addition of table counts in cols 3-6 and rows 2-7 only.

2) *perturb_ out.txt*
[stored in folder Perturb_v3]

For each table processed, this file reports whether the table has consistent marginals, or had inconsistent marginals requiring pre-perturbation correction.

e.g.

```
=== table02 - has consistent marginals
```

3) <*tablename*>*_0.fmt*
[stored in folder *Perturb_v3*]

A copy of the original SAS table counts, as input from <*tablename*>*_raw.fmt*, with inconsistent table marginals adjusted to fit internal table counts

4) <*tablename*>*_1.fmt*
[stored in folder *Perturb_v3/Outputs*]

A set of the original above-threshold table counts perturbed by random rounding to an adjacent multiple of 5.

5) <*tablename*>*_2.fmt*
[stored in folder *Perturb_v3/Outputs*]

A set of the original above-threshold table counts perturbed by random rounding to an adjacent multiple of 3.

6) <*tablename*>*_3.fmt*
[stored in folder *Perturb_v3/Outputs*]

A set of the original above-threshold table counts perturbed by random rounding to an adjacent multiple of 3, implemented only for counts < 3.

7) <*tablename*>*_4.fmt*
[stored in folder *Perturb_v3/Outputs*]

A set of the original above-threshold table counts perturbed by Barnardisation (random addition of +/- 1 or 0). (prob of change = 2 x 0.1)

8) <*tablename*>*_5.fmt*
[stored in folder *Perturb_v3/Outputs*]

A set of the original above-threshold table counts perturbed by Barnardisation (random addition of +/- 1 or 0). (prob of change = 2 x 0.02)

Create_Aggregates_v2

[in folder *Create_Aggregates_v2*]

The main purpose of *Create_Aggregates* is to produce data for analysis by *SDC_Direct_Impacts* and *SDC_Indirect_Impacts*. To this end, *Create_Aggregates* reads in tabular data for a set of user-supplied areas and performs one or more of the following operations:

- Selects a user-defined no. of samples from the set of user-supplied areas
- Stratifies the user-supplied data (if required); if strata are required, the same user-specified no. of samples are selected from each strata.
- Outputs into one file all samples for a given combination of table name, strata type, sample type, sample size and method of disclosure control
- Calculates and outputs a series of user-defined percentages based on the sampled count data (if required)

Each sample consists of 1 or more user-supplied areas. For requested sample sizes > 1, each sample is based on aggregation of an appropriate number of user-supplied areas.

Sampling is undertaken using one of a number of alternative sampling strategies (further details below): each area in turn; sequential sets of areas; random areas or sets of areas (with replacement); geographically clustered areas.

Program limits

No. of tables: 20
No. of table columns: 40 [includes marginals]
No. of table rows: 40 [includes marginals]
No. of unique contributing cell types per table: 50 (the number of cells contributing to a table cell count will range from 1 [internal cell], to no. of cells contributing to the value of the table total) [for more details, see documentation of *SDC_Direct_Impacts*)
No. of percentages: 100
No. of samples: Number of user-defined samples to be taken from user-supplied area data (Minimum: 1; Maximum: 1000)
No. of user-supplied areas: 5500

Program inputs

1) Unperturbed tabular data
[stored in folder *Perturb_v3/Outputs*]

For each table to be assessed a file containing pre-perturbation counts (space or comma-separated counts) is required, including a count of the population at risk of appearing in the table; with counts for successive areas preceded by a line containing an appropriate area identifier. The area identifier can be any length, and include any keyboard character other than quote marks. It is recommended that data are set out in tabular form (i.e. in rows and columns), but vector input is permissible.

For example:

```
01ABFQ19
  340  0  320  20  140  10
  30  0  30  0  15  0
01ABFQ20
  290  0  275  15  110  5
  30  0  25  0  15  0
01ABFQ21
```

```
360  0  355  5  145  5
40  0  40  0  15  0
```

File naming should follow the convention <*table name*><*variant name*>*.fmt* . For example, the original versions of SAS Table 08 (*s08*) could be stored in the file *s08_v0.fmt* or *table8_original.fmt*. [The underscore is required as a prefix to the variant name]

2) Perturbed tabular data

For each table to be assessed, a separate file containing post-perturbation counts is required for each variant of statistical disclosure control to be assessed. The layout of these files should conform to that described for unperturbed tabular data above. File naming should follow the convention <*table name*><*variant name*>*.fmt* . For example, a perturbed version of SAS Table 08 could be stored in the file *s08_v1.fmt* or *table8_round_base3.fmt*. [The underscore is required as a prefix to the variant name]

3) Table maps

For each table to be assessed, a separate file detailing the number of table rows and columns (including marginals). File naming should follow the convention <*table name*>*.map* .

For example the structure (mapping) of the following table

SAS Table 06(a)

Sex and Age	Total Persons	**Ethnic group**										Persons born in Ireland
		White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups		
										Asian	Other	
Total Persons	115	94	4	0	0	3	0	0	12	0	2	7

would be stored in a file with a name such as *s06.map* and would contain the following line:

```
1  11
```

[i.e. 1 row; 11 columns]

Those intending to use the *SDC_Direct_Impacts* part of the **SDC-i** suite should note that *Create_Aggregates* can also read in and process the more complicated table mappings required to obtain the maximum range of disclosure impact measures from *SCD_Direct_Impacts* (see p.18), obviating the need to have multiple versions of each table map.

3) *Aggregates input tables.txt*
[stored in folder *Create_Aggregates_v2*]

A file containing a list of the tables for which the impact of statistical disclosure control is to be evaluated, one table name per line. The names given for each table should match the *table name*s used in (1) and (2) above.

```
s06
s12
s20
```

4) *sdc_variants.txt*
[stored in folder *Create_Aggregates_v2*]

This file should contain a list of the sdc variants to be sampled/stratified/aggregated (each variant name in quotes). The *variant names* listed should match those used to name the input files of perturbed and unperturbed data.

E.g.

```
_v0
_v1
_v2
_v3
```

5) *thresholds.txt*
[stored in folder *Create_Aggregates_v2*]

A file listing the minimum no. of residents and households required for an area to be above threshold (space- or comma-separated).

Eg.

```
100 40
```

6) *threshold.fmt*
[stored in folder *SAS_Inputs*]

A file listing the number of private households and residents present in each geographic area for which input table(s) are supplied.

In the file, one area is described per line. For each area three pieces of information are given, separated by commas:
(1) Area name – maximum of 8 characters, in quote marks
(2) Number of private households containing residents
(3) Number of residents in private households

E.g.

```
"BYFA01",           38,          113
"BYFA02",          124,          259
"BYFA03",          221,          413
"BYFA04",           46,          210
"BYFA05",          109,          351
"BYFA06",           93,          230
"BYFA07",            0,            0
```

Note that this file provides as a master list against which *Create_Aggregates* checks all other area-level data inputs. Incorrectly ordered or missing area-level data in other files will cause program execution to halt, and an error message to be generated.

7) *Create_Aggregates_percentages.map*
[stored in folder *Create_Aggregates_v2*]

End-users often use percentages calculated from perturbed data as part of their analyses. *Create_Aggregates* calculates these percentages for each sample on the fly. This is necessary to allow for the cumulative impact of aggregating areas with different denominators. This file does

not have to be supplied if the *use counts/percentages flag* in *Create_Aggregates_run_parameters.txt* (see below) is set to 0.

First row of file = output format for calculated percentages (no. of rows; no. of cols) [currently read in but not used]

Each subsequent row:
One row per percentage, listing in comma-separated order, with text in quotes:
(i) Percentage name
(ii) No. of numerator components
(iii) Name of table from which numerator component is drawn (e.g. table12")
(iv) Column (x) within table from which numerator component is drawn
(v) Row (y) within table from which numerator component is drawn
(vi) Operator ["+" or " –"] to be used to sum current and next numerator component (if there is a next numerator component)
(vii) Repeat of (iii) to (vi) for each numerator component
(viii) No. of denominator components
(ix) Name of table from which denominator component is drawn (e.g. "table12")
(x) Column (x) within table from which denominator component is drawn
(xi) Row (y) within table from which denominator component is drawn
(xii) Operator ["+" or " –"] to be used to sum current and next denominator component (if there is a next denominator component)
(xiii) Repeat of (ix) to (xii) for each denominator component

E.g.

"pltill", 1, "table12", 1, 1, 1, "table35", 1, 1

The percentage "pltill" has one numerator component, drawn from "Table12", column 1, row 1; and one denominator component, drawn from "Table35", column 1, row 1.

"pownocc", 2, "s20", 2, 1, "+", "s20", 3, 1, 1, "s20", 1, 1

The percentage "pownocc" has two numerator components. The first is drawn from table "s20", column 2, row 1; and should be added to the second numerator component, drawn from table "s20", column 3, row 1. There is one denominator component, drawn from table "s20", column 1, row 1.

8) *Strata source file*
[Stored in folder SAS_Inputs]

If the creation of stratified samples is required, *Create_Aggregates* needs as an additional input a file providing area-level measures, for all user-supplied areas, of the stratification variable to be used. The name of this strata source file is specified as a user-supplied parameter in the file *Create_Aggregates_run_parameters.txt*. This file does not have to be supplied if the *sampling strata* flag in *Create_Aggregates_run_parameters.txt* (see below) is set to zero.

Each line of this file should contain an area name (in quotes), separated by a comma from the associated area-level measure. The first line of the file should contain header IDs for each data column, each ID in quote marks, and again comma-separated.

E.g.

```
"Zone ID","popdens"
"04BXFA01",5725
"04BXFA02",14000
"04BXFA03",11766.66667
"04BXFA04",1592.857143
"04BXFA05",6966.666667
```

The header for the second column will be used by *Create_Aggregates* outputs to identify the stratification variable in use.

9) *Create_Aggregates_run_parameters.txt*
[stored in folder *Create_Aggregates_v2*]

List of user-defined run parameters:

```
"No. of samples:", 100
"Sampling strata [1=All;2=P20/P80;3=All/P20/P80]:", 2
"Sample type:", 2
"Sample size:", 20
"Report thresholding of areas [on/off]:", 0
"Report sorting of areas [on/off]:", 0
"Report sample membership [on/off]:", 1
"Report table mapping [on/off]:", 1
"Report percentage mapping [on/off]:", 1
"Use counts/percentages [0=count;1=%; 2=count & %]:", 1
"Strata source file:", "popdens.fmt"

=====================================================
For all on/off switches, 1 = on; any other number = off
```

No. of samples: user-defined no. of samples to be taken from the set of user-supplied areas (minimum: 1; maximum: 1000)

Sampling strata: identifies range of user-supplied areas from which samples should be taken. 1: whole dataset; 2: top/bottom quintiles of a user-supplied area-level measure (specified via *strata source file* – see below); 3: whole dataset/top quintile/bottom quintile.

Sample type [sampling strategy]:

(1) Select every user-supplied area in turn (maximum 1000 areas) [A]. N.B. To process every user-supplied area in turn, sampling strata and sample size (see below) must both be set to 1.

(2) Select consecutive areas, but starting from random starting point [C]. This allows for some simulation of spatial contiguity if consecutive areas are spatial neighbours and sample stratification is not turned on; or can help maximise sample homogeneity if used in conjunction with sampling strata (consecutively ranked areas within strata will be selected).

(3) Select areas at random (with replacement) [R]

(4) Select areas within geographical cluster [G] (This only works for datafiles containing clusters of 50 above threshold areas, provided sample stratification is set to 1, and is not really intended for public use in current form)

Sample size: Each sample consists of a user-defined number of user-supplied areas. There is no maximum sample size, but for sensible results the following guidelines are recommended:-

Sampling type 1: sample size = 1 (sample stratification = 1)

Sampling type 2: max. recommended sample size = no. of user-supplied areas/20, if sampling strata turned off; max. sample size = no. of user-specified areas / (100) if sampling strata turned on)
Sampling type 3: max. recommended sample size as per sampling type 2.
Sampling type 4: max. recommended sample size = cluster size (sample stratification = 1)

Report thresholds of areas: lists user-supplied areas contained in *threshold.fmt*, identifying those that fall below threshold criteria contained in *thresholds.txt*. Results placed in *Create_Aggregates_temp.txt*

Report sorting of areas: Lists ranking of all above-threshold areas given user-supplied stratification variable (c.f. *strata source file* parameter below) in the file *Create_Aggregates_temp.txt*

Report sample membership: writes list of user-supplied areas constituting each sample to *Create_Aggregates_temp.txt* [warning: list will be very long if user-specified no. of samples and sample size are large]

Report table mapping: Provides visualisation of supplied table mapping in *Create_Aggregates_temp.txt* (if requested).

Report percentage mapping: if this flag is turned on and *Use counts/percentages* (below) is set to 1 or 2, then reports table, row and column of each cell contributing to calculation of percentage numerator and denominator to *Create_Aggregates_temp.txt*.

Use counts/percentages: create 'aggregated' files of based on: 0 - counts only; 1 - percentages only; 2 - counts and percentages. [N.B. Percentages can only be calculated if data files for all tables listed in *Create_Aggregates_percentages.map* are supplied]

Strata source file: name of file containing area-level data, on basis of which areas are to be stratified. Name of file should be in quote marks, as in the example above, and should be given relative to *SAS_Inputs* folder.

Program outputs

1) *agg_output_tables_list.txt*
[stored in folder *Create_Aggregates_v2*]

The name of each aggregated file produced by *Create_Aggregates* is written to this file, including the full pathname.

E.g.

```
d:\research\disclosure control\Aggregates\table02_v0_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table06a_v0_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table02_v1_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table06a_v1_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table02_v2_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table06a_v2_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table02_v3_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table06a_v3_P20[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table02_v0_P80[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table06a_v0_P80[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table02_v1_P80[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table06a_v1_P80[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table02_v2_P80[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table06a_v2_P80[popdens]_n20[C]_s100.fmt
d:\research\disclosure control\Aggregates\table02_v3_P80[popdens]_n20[C]_s100.fmt
```

```
d:\research\disclosure control\Aggregates\table06a_v3_P80[popdens]_n20[C]_s100.fmt
```

2) *Create_Aggregates_temp.txt*
[stored in folder *Create_Aggregates_v2*]

Scratch output file used to store program progress reports, and information such as list of input/above-threshold areas, percentage mappings etc., as requested by user via *Create_Aggregates_run_parameters.txt*.

3) Aggregated sample files
[Stored in folder Create_Aggregates_v2/Output]

When running *Create_Aggregates*, users can specify a wide variety of options, as listed below. For each unique combination of options, one output is created, containing data for a set of sample areas. The options selected determine the name of the output file.

User options	***Output naming conventions***
Base table name	*table name* / **percentages**
Disclosure control variant	*variant name*
Sampling strata	**P20** or **P80** or **All**
Strata type	[*strata_name*]
No. of areas per sample	n*X*
Sampling strategy	**[A]** or **[C]** or **[R]** or **[G]**
No. of samples	s*X*
Strata data source file	*strata_data_source_file*

For example

```
S06a_v0_P20[Popdens]_n20[R]_s1000.fmt
```

Would be a file containing 1000 samples, based on 20 (aggregated) EDs per sample, drawn randomly from the lower quintile of the user-supplied *PopDens* distribution, using unperturbed counts (sdc variant 0) from SAS Table 06(a).
An example of the contents of this output file is:

```
s06a_v0_P20[popdens]_n20[R]_s1
 9834  7351  371  180  100  687  212  666  50  92  125  328
 4807  3547  175  84  45  335  122  360  21  49  69  145
 5027  3804  196  96  55  352  90  306  29  43  56  183
s06a_v0_P20[popdens]_n20[R]_s2
 9780  8011  461  258  137  417  110  60  64  130  132  215
 4629  3782  201  125  62  217  52  30  34  59  67  96
 5151  4229  260  133  75  200  58  30  30  71  65  119
.
.
.
s06a_v0_P20[popdens]_n20[R]_s1000
 9972  8276  477  362  91  156  34  346  37  98  95  229
 4747  3926  213  174  43  84  16  186  16  46  43  102
 5225  4350  264  188  48  72  18  160  21  52  52  127
```

Note that the header to each aggregated sample in the file follows the same naming convention as the file itself, with the exception that the number following "_s" denotes the actual sample number, rather than the total number of samples in the file.

SDC_Direct_Impacts

SDC_Direct_Impacts measures the direct impact of disclosure control measures on tabular outputs.

A typical tabular output comprises both interior and marginal counts. In this guide:
- A *marginal* is any table cell whose value, prior to the application of disclosure control measures, equals the sum of two or more *counts* present elsewhere in the same table.
- A *count* is any table cell that is not a *marginal*.

The main input to *SDC_Direct_Impacts* is a set of pre- and post-perturbation table counts and marginals (and/or percentages based upon these counts).

The main output is a set of statistics summarising the difference between the pre- and post-perturbation table counts and/or percentages. These outputs include a range of cellular and tabular measures, as well as an optional assessment of differences in pre- and post-adjustment area rankings.

SDC_Direct_Impacts can also summarise the average impact of disclosure control across multiple table layouts (e.g. tables with: differing numbers of counts; focus on more or less rare population sub-groups; marginals based on summation across differing numbers of cells).

SDC_Direct_Impacts, if used in conjunction with the outputs from *Create_Aggregates*, is also capable of summarising the average impact of disclosure control across multiple versions of the same table generated by alternative sampling strategies (e.g. inputs based upon differing sized aggregates of input areas; inputs drawn from different strata, such as urban vs. rural or 'rich' vs. 'poor').

SDC_Direct_Impacts optionally allows for assessment of the impact of 'indirect perturbation'. Indirect perturbation occurs when a table marginal is derived from summation of perturbed table counts, rather than from direct perturbation of the original marginal count, even if the original input marginal counts were independently perturbed.

Program limits

Input tables:	20
Samples/areas per table:	1000
Rows / columns/ cells per table:	40 / 20 /800
Total cells in all tables:	16000
Cell types[1] (count + marginal(s)) per table):	50
Cell types[1] across all input tables:	200
Marginal mappings per table:	30

[1] A cell's 'type' is defined by the number of counts upon which its original value depends. 'Cell types' is the number of unique cell types in an input table/dataset (including interior cell counts of type '1').

Program Run time

Increases with both the number of measures of fit requested and the number of pre/post adjustment cell counts to be evaluated. Using the default settings with the supplied benchmark data (11,410 cell counts) program run-time is 4 minutes on a Pentium IV 3GHz desktop PC with 0.5Gb RAM. Execution speed will slow dramatically if adequate RAM is not provided.

PROGRAM INPUTS

1) Program pathnames

(a) Program path

If running *SDC_Direct_Impacts* direct from its compiled executable version, the root folder (Program path) is automatically assigned as the folder in which the executable code is located.

If compiling and running *SDC_Direct_Impacts* via *VisualBasic* change the line of code

```
ProgramPath = "C:\Temp\Test SDCi"
```

to point to the folder a root folder of your own choice (e.g. "C:\Program Files\SDCi"). Note that this pathname should NOT end with a slash.

Alternatively, to compile and run the code as an executable, comment out the above line of code, and comment in the preceding line: `ProgramPath = CurDir()`

(b) Input_and_output_paths.txt

SDC_Direct_Impacts requires a number of data inputs. To allow maximum flexibility, users are able to specify the locations for four types of input data:

InputCounts: Pre- and post-adjustment cell counts to be compared
TableMappings: Table mappings describing layout of each input table (required)
StrataData: Data to be used for creation of stratified samples (optional)
RunParameters: Files containing program run-time parameters (required)

The file *input_and_output_paths.txt* lists these input/output sources, each followed by a pathname, defined relative to the program execution root folder, pointing to the relevant user-specified folder:

```
"StrataDataPath", "\Strata Data\"
"TableMappingsPath", "\Table mappings\"
"RunParametersPath", "\Parameters\"
"InputCountsPath", "\SDCi Input Counts\"
```

Note that, if modifying the default settings above, the quote marks, comma, and the first and final backward slash at the start and end of each pathname should all be retained.

2) Pre-perturbation counts

[Stored in the *InputCounts* folder pointed to in *Input_and_output_paths.txt*]

One file per table, containing the original table counts, prior to the application of statistical disclosure control, for 1 – 1000 areas/samples. (A sample = 1 or more areas previously selected at random, and aggregated if appropriate, from a larger set of user-supplied areas). These files may be supplied by the user, or produced using *Create_Aggregates*.

Files supplied directly by the user should use the following naming convention:

<*table name*>**_v***n*.**fmt**

where *n* is any user-specified number indicating a particular disclosure control variant. It is

recommended, but not essential, that 0 is used to indicate files containing the original unperturbed counts.

E.g. *User_supplied_table_**v**0.**fmt***

Within each file, it is recommended that counts are laid out in rows and tables as per the published version, although supply of counts in vector format is also supported.

The counts (including marginals) should be space or comma separated (no commas at ends of rows).

For example, the table

SAS Table 06 Ethnic group of Residents by Age and by Sex

Enumeration District: BYFA01

Sex and Age	Total Persons	Ethnic group										Persons born in Ireland
		White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups		
										Asian	Other	
Total Persons	115	94	4	0	0	3	0	0	12	0	2	7
Males	54	45	1	0	0	0	0	0	6	0	2	1
Females	61	49	3	0	0	3	0	0	6	0	0	6
0-4	6	5	0	0	0	0	0	0	1	0	0	0
5-15	5	3	0	0	0	0	0	0	2	0	0	0
16-29	52	44	1	0	0	0	0	0	5	0	2	5
30<pa	42	36	2	0	0	3	0	0	1	0	0	0
Pa and over	9	5	0	0	0	1	0	0	3	0	0	2

would be represented in the file *s06_v0.fmt* as

```
s06_v0_s1.fmt
  115   94    4    0    0    3    0    0   12    0    2    7
   54   45    1    0    0    0    0    0    6    0    2    1
   61   49    3    0    0    3    0    0    6    0    0    6
    6    5    0    0    0    0    0    0    1    0    0    0
    5    3    0    0    0    0    0    0    2    0    0    0
   52   44    1    0    0    0    0    0    5    0    2    5
   42   36    2    0    0    3    0    0    1    0    0    0
    9    5    0    0    0    1    0    0    3    0    0    2
```

Or

```
s06a_v0_s1.fmt
  115,  94,   4,   0,   0,   3,   0,   0,  12,   0,   2,   7
   54,  45,   1,   0,   0,   0,   0,   0,   6,   0,   2,   1
   61,  49,   3,   0,   0,   3,   0,   0,   6,   0,   0,   6
    6,   5,   0,   0,   0,   0,   0,   0,   1,   0,   0,   0
    5,   3,   0,   0,   0,   0,   0,   0,   2,   0,   0,   0
   52,  44,   1,   0,   0,   0,   0,   0,   5,   0,   2,   5
   42,  36,   2,   0,   0,   3,   0,   0,   1,   0,   0,   0
    9,   5,   0,   0,   0,   1,   0,   0,   3,   0,   0,   2
```

As shown above, the counts for each area must be preceded by a header. This header should be used to identify the area which the set of counts represents in a way which is meaningful to the user, and should be in quotes if the identifier includes a space.

Data for the next area should start on the next empty row. (Do NOT leave a blank row between areas.) For example:

```
s71_v0_s1
 7399  104  7226  69   2991   40
  718    9   709   0    298    0
s71_v0_s2
 7021  121  6823  77   3057   43
 706    12   694   0    307    0
```

Files created via *Create_Aggregates* automatically conform to the above requirements.

3) Post-perturbation counts

[Stored in the *InputCounts* folder pointed to in *Input_and_output_paths.txt*]

One file per table variant, containing the perturbed table counts arising from a particular disclosure control method, for 1 – 1000 areas/samples. (A sample = 1 or more areas previously selected at random, and aggregated if appropriate, from a larger set of user-supplied areas). Files containing perturbed counts for a set of samples may be supplied by the user themselves, or produced using *Create_Aggregates*. Users lacking perturbed counts may produce perturbed versions of user-supplied counts using *Perturb*.

Input files supplied directly by the user should use the following naming convention:

<*table name*>_v*n*.fmt

where *n* is any user-specified number indicating a particular disclosure control variant.

E.g. *User_supplied_table_v2.fmt*

It is recommended, but not essential, that 0 is reserved to indicate files containing the original unperturbed counts.

The names of input files created via *Create_Aggregates* should be left unchanged.

For example, the following three files would contain the perturbed counts arising from three different statistical disclosure control methods:

```
S06_v1.fmt
S06_v2.fmt
S06_v3.fmt
```

The file layout required is the same as that used for original counts, as outline in (2) above.

4) Table mappings

[Stored in the *TableMappings* folder pointed to in *Input_and_output_paths.txt*]

For each input table, a file is required specifying the table structure (rows/columns/marginals etc.). For this file the naming convention <*table name*>*.map* should be followed (e.g. *User_supplied_table.map* or *s06.map* for the examples presented in (2) above).

Creating an appropriate table mapping is by far the most onerous part of preparing data for input to *SDC_Direct_Impacts* (and to *Perturb*). Full details on how to create such table mappings are set out below, but in general the file will include: (i) number of rows and columns in table; (ii) row counts which sum to give row marginal(s) [if any]; (iii) column counts which sum to give column marginal(s) [if any]

<u>Example 1</u>: **Table containing only independently perturbed table counts**

Sex and Age	**Ethnic group**										Persons born in Ireland
	White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups: Asian	Other groups: Other	
Total Persons	94	4	0	0	3	0	0	12	0	2	7

Given that all of the counts in the above table are independent of each other, the full description of this table required by *SDC_Direct_Impacts* is:

```
1 11
```

Description Row 1: number of rows in table, followed by number of columns (above example = table with 1 row and 11 columns)

Example 2: Table containing one dependent table marginal

Sex and Age	Total Persons	**Ethnic group**										Persons born in Ireland
		White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups		
										Asian	Other	
Total Persons	115	94	4	0	0	3	0	0	12	0	2	7

The original 'total persons' count in the above table is based on the sum of the interior ethnic group counts. Additional information is required, therefore, mapping the contribution of each table count to this table marginal.

In this case the full table description required by *SDC_Direct_Impacts* would be:

```
1 12
1 -1 2 3 4 5 6 7 8 9 10 11 0
```

The description is compiled as follows:

Description Row 1: number of rows in table, followed by number of columns (above example = table with 1 row and 12 columns)

Description Row 2, first number: flag to indicate whether following numbers give a mapping for a row or column marginal [1 = row, 2 = column]. In this case 'total persons' is a row marginal (sum of counts in row), so first number in row 2 of the table mapping is 1.

Second row, remaining numbers: A flag is given for each column in the table, reading from left to right, as follows:

Flag	Meaning
-1	Column containing the row marginal being mapped
>0	Column containing a count that contributes to the row marginal being mapped
0	Column containing a count that does NOT contribute to the row marginal being mapped

When appropriate, the same flags are used to record the contribution of each row to a column marginal (reading from top to bottom).

In the above example, the row marginal recorded in column 1 [column 1 flagged with a –1] is the sum of columns 2 through 11 [each column flagged by a positive number]. Column 12 is present only due to table concatenation and does not contribute to the calculation of the table marginal. It is therefore flagged with a 0.

Example 3: Table with dependent column and row marginals

Sex and Age	Total Persons	Ethnic group										Persons born in Ireland
		White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups		
										Asian	Other	
Total Persons	115	94	4	0	0	3	0	0	12	0	2	7
0-4	6	5	0	0	0	0	0	0	1	0	0	0
5-15	5	3	0	0	0	0	0	0	2	0	0	0
16-29	52	44	1	0	0	0	0	0	5	0	2	5
30<pa	42	36	2	0	0	3	0	0	1	0	0	0
Pa and over	9	5	0	0	0	1	0	0	3	0	0	2

In the above table the original 'total persons' counts in each row and column are based upon the sum of various interior counts. Additional information is required to 'map' the contribution of table counts to each column and row table marginal.

In this case the appropriate table description would be:

```
6 12
1 -1 2 3 4 5 6 7 8 9 10 11 0
2 -1 2 3 0 0 0
```

Description Row 1: 6 rows by 12 columns

Description Row 2: Row mapping (first number =1); column 1 is a row marginal [-1]; columns 2 through 11 sum to give total in column 1 [values >0]; 12th column does not contribute to row marginal [0]

Description Row 3: Column mapping (first number=2); row 1 is a column marginal [-1]; rows 2 and 3 sum to give total in column 1 [values > 0]

Example 4: Table with multiple dependent row and column table marginals

This final example is based upon a complex table containing multiple totals and sub-totals (see next page). Given that all table marginals are based on the sum of the relevant interior counts to be found in the body of the table, this table requires mappings for one row marginal and six column marginals:

```
28 11
1 -1 2 3 4 5 6 7 8 9 10 0
2  0 -1  3  4  5  6  7  8  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2  0  0  0  0  0  0  0  0  0 -1 11 12 13 14  0  0  0  0  0  0  0  0  0  0  0  0
2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1 17 18 19 20 21 22  0  0  0  0
2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1 25 26 27 28
2 -1  2  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1 16  0  0  0  0  0  0  0 24  0  0  0
```

Note the need for one mapping per table marginal being mapped.

Note also that, in this example, to save time, some table marginals are expressed as the sum of other table marginals

Table 08 Economic position: residents aged 16 and over

Sex by economic position	Total aged 16 and over	Age									Students (Econ. active or inactive)
		16-19	20-24	25-29	30-34	35-44	45-54	55-59	60-64	65+	
Males											
Economically active											
Employees full-time											
Employees part-time											
Self-emp. + employees											
Self emp. 0 employees											
On a govt. scheme											
Unemployed											
Student (incl. Above)											
Economically inactive											
Students											
Permanently sick											
Retired											
Other inactive											
Females											
Economically active											
Employees full-time											
Employees part-time											
Self-emp. + employees											
Self emp. 0 employees											
On a govt. scheme											
Unemployed											
Student (incl. Above)											
Economically inactive											
Students											
Permanently sick											
Retired											
Other inactive											

5) *SDC_Direct_Impacts_run_parameters.txt*

[Stored in the *RunParamters* folder pointed to in *Input_and_output_paths.txt*]

The main purpose of *SDC_Direct_Impacts* is to evaluate the difference between perturbed and unperturbed count and percentage data. Users can select from a wide variety of goodness-of-fit measures at cellular, tabular and cross-table (i.e. global average) measures by modifying the relevant options in the file *SDC_Direct_Impacts_run_parameters.txt*. Options should be registered by changing the relevant values to the right of the comma on each line. The default settings are shown below. Please note that the spacing (blank lines) between sections is vital to the correct execution of the program, and should not be altered in any way.

Following the example file, the remainder of this section explains the meaning of the various parameters and the options available for each.

```
"=== file information on input counts ==="

"Data source [Create_Aggregates/User]:             ", "Create_Aggregates"
"No. of samples:                                   ", 10
"Sampling strata [1=All;2=P20/P80;3=All/P20/P80]:  ",  2
"Sample type:                                      ",  3
"Sample size:                                      ", 20
"Report table mapping [on/off]:                    ",  1
"Use counts/percentages [0=count;1=%; 2=count & %]:",  0
"Strata source file:", "popdens.fmt"

"=== Report types ==="

"Table Totals [on/off]:              ", 0

"Table-specific, Area-specific, Cell-based [on/off]:   ", 0
"Table-specific, Area-specific, Table-based [on/off]: ", 0
"Table-specific, Cross-area,      Cell-based [on/off]:   ", 0
"Table-specific, Cross-area,     Table-based [on/off]:  ", 0
"Cross-table,     Area-specific,  Table-based [on/off]: ", 0
"Cross-table,     Cross-area,      Table-based [on/off]:  ", 1

"Correct Rank [on/off]:              ", 1
"Correct Class [on/off]:             ", 1
"Correct/Neighbouring Class [on/off]: ", 1

"=== Cell-based measures of fit ==="

"cell_exp [on/off]:     ", 0
"cell_obs [on/off]:     ", 0
"cell_changed [on/off]: ", 0
"cell_TE [on/off]:      ", 0
"cell_Z [on/off]:       ", 0
"cell_NFC [on/off]:     ", 0
"cell_Zm [on/off]:      ", 0
"cell_NFCm [on/off]:    ", 0

"Cell_Summary, Max [on/off]:      ", 1
"Cell_Summary, 95%-tile [on/off]: ", 1
"Cell_Summary, mean [on/off]:     ", 1
"Cell_Summary, 5%-tile [on/off]:  ", 1
"Cell_Summary,min [on/off]:       ", 1

"=== Table-based measures of fit ==="

"Table_frequency (of cell type) [on/off]: ", 1
"Table_n_changed [on/off]:                ", 1
"Table_p_changed [on/off]:                ", 1
"Table_max_change [on/off]:               ", 1
"Table_maxPchange [on/off]:               ", 1
"Table_TotalError [on/off]:               ", 1
"Table_TAE [on/off]:                      ", 1
"Table_RAE [on/off]:                      ", 1
"Table_SAE [on/off]:                      ", 1
"Table_Sq_Error [on/off]:                 ", 1
"Table_RMSE [on/off]:                     ", 1
"Table_SSZ [on/off]:                      ", 1
"Table_NFC [on/off]:                      ", 1
"Table_NFT [on/off]:                      ", 1
"Table_SSZm [on/off]:                     ", 1
"Table_NFCm [on/off]:                     ", 1
"Table_NFTm [on/off]:                     ", 1
"Table_Gibsons_D [on/off]:                ", 1
"Table_Cramers_V [on/off]:                ", 1
"Table_PearsonsR [on/off]:                ", 1
"Table_ChiSquare [on/off]:                ", 1
"Table_TVCC [on/off]:                     ", 1
"Table_v_expcells [on/off]:               ", 1
"Table_v_obscells [on/off]:               ", 1

"Table_Summary, Max [on/off]:             ", 1
"Table_Summary, 95%-tile [on/off]:        ", 1
"Table_Summary, mean [on/off]:            ", 1
"Table_Summary, 5%-tile [on/off]:         ", 1
"Table_Summary, min [on/off]:             ", 1


=================================================================
Note 1. For all on/off switches, 1 = on; any other number = off
```

5(a) Information on input counts

Data source [Create_Aggregates/User]: For user-supplied inputs, set option to *User*. If the program *Create_Aggregates* has been used to create the input files of perturbed/unperturbed counts, set to *Create_Aggreagtes*.

No. of samples: No. of input areas (i.e. no. of areas for which data are supplied via the input files described in (1) and (2) above).

Sampling strata [1=All;2=P20/P80;3=All/P20/P80]: If the data source is "User", then sampling strata may be set to any whole number as the actual value chosen will have no impact on program operation; if the source is "Create_Aggregates", strata selection should reflect that previously used in *Create_Aggregates*.

Sample type: If the data source is "User", then sample type should be set to any whole number, as the actual value chosen will have no impact on program operation; if the source is "Create_Aggregates", sample type should reflect that used in *Create_Aggregates*.

Sample size: If the data source is "User", then sample type should be set to any whole number, as the actual value chosen will have no impact on program operation; if the source is "Create_Aggregates", sample size should reflect that used in *Create_Aggregates*.

Report table mapping: If set to 1, the output file *SDC_Direct_Impacts_results.txt* (located in the *ProgramPath* folder) will contain a table mapping indicating, for each table cell, the number of other table cells on which its value depends. This is useful for checking that table mappings have been properly declared. If set to 0, table mappings will not be reported.

Use counts/percentages [0=count;1=%; 2=count & %]: A choice of whether assessment of disclosure control impact should be made for counts only [0]; percentages only [1]; or both counts and percentages [2]. Note that options [1] and [2] require the user to supply percentage mappings (see (8) below).

Strata source file: If the *sampling_strata* option has been set to [2] or [3], the name of the datafile upon which stratification by *Create_Aggregates* was based should be specified (e.g. "popdens.fmt"); else leave set to the default "None".

5(b) Report types

The output from *SDC_Direct_Impacts* is written to the file *SDC_Direct_Impacts_results.txt*, located in the *ProgramPath* folder. In addition to the cell-based and table-based measures chosen (see (c) and (d) below), the precise contents of this file depends upon the report-type selected. The basic report types available are outlined below. For all report types, a parameter value of 0='off', 1='on'.

5b(i) Table Totals: For some input tables, the sum of the internal cell counts contributing to the overall table total may not equal the actual table total. If required, both table totals will be reported, for both the original and perturbed table variants. For example:

```
=== Revised table totals for s06a ===

Table s06a As published          : Expected total  9834  Observed total  9831
Table s06a Sum of internal counts: Expected total  9834  Observed total  9882
```

5b(ii) Table-specific, Area-specific, Cell-based: reports all user-requested cell-based measures for each table cell, in each input table, for each input area. The available cell-based measures are listed in the section headed 'cell-based measures' below.

The following example report includes three of the available cell-based measures:

```
=== Table-specific, Area-specific, Cell-based report for s06a (Sample     1) ===

cell_exp
    9834    7351     371     180     100     687     212     666      50      92     125     328
    4807    3547     175      84      45     335     122     360      21      49      69     145
    5027    3804     196      96      55     352      90     306      29      43      56     183

cell_changed
       1       1       1       1       1       1       1       1       1       1       1       1
       1       1       1       1       0       1       1       0       1       1       1       1
       1       0       1       0       1       1       1       0       1       1       1       1

cell_diff
      -3      -1       1      -3       2      -3      -2       3      -5      -5      -2       2
      -4      -1      -1       6       0       4       4       0       6       5       9      -4
      -5       0      -4       0      -4       5      -3       0      13       8       1      -3

=== Table-specific, Area-specific, Cell-based report for s06a (Sample     2) ===

cell_exp
    9780    8011     461     258     137     417     110      60      64     130     132     215
    4629    3782     201     125      62     217      52      30      34      59      67      96
    5151    4229     260     133      75     200      58      30      30      71      65     119

cell_changed
       1       1       1       1       1       0       1       1       1       1       1       1
       1       1       1       1       1       1       1       0       1       1       1       0
       1       1       1       1       1       1       1       1       1       1       1       1

cell_diff
      -3      -7       1      -6      -2       0      -2       3       8      -1       3      10
      -3      -2      -3      -2       1      -4       2       0       5       4      -7       0
      -6      -5     -11      -4       3       4       2       3      -3       7      -2      10

=== Table-specific, Area-specific, Cell-based report for s06a (Sample     3) ===

etc…
```

As may be seen from above, all requested cell-based measures are reported for each input area (sample) in turn. The layout of the cells directly mirrors the layout of the cells as input to *SDC_Direct_Impacts*, with the number of columns and rows conforming to that recorded in the table mapping. The example above presents results for the following input table layout:

Sex and Age	Total Persons	**Ethnic group**										Persons born in Ireland
		White	Black C'bean	Black African	Black other	Indian	P'stani	B'deshi	Chinese	Other groups: Asian	Other groups: Other	
Total Persons	115	94	4	0	0	3	0	0	12	0	2	7
0-4	6	5	0	0	0	0	0	0	1	0	0	0
5-15	5	3	0	0	0	0	0	0	2	0	0	0

WARNING: for large input datasets, with many areas and/or many tables, the potential size of the output file produced by this report option is very large. The main purpose of this reporting option is simply to aid quality assurance of outputs from *SDC_Direct_Impacts* using small pilot datasets.

5b(iii) Table-specific, Area-specific, Table-based: reports all user-requested table-based measures for each user-supplied input table, for each input area (sample). The available table-based measures of fit are described below in the section 5(d) headed 'table-based measures'.

For example, if the number of cells changed by disclosure control (*n_changed*) is requested, the resulting output would look like:

```
=== Table-specific, Area-specific, Table-based report for s06a ===

                  Cell type (no. of contributing cells count depends upon)
Sample  Measure       Marginal     Internal    All         1           2           10          20
     1  n_changed      14.000000   17.000000   31.000000   17.000000   11.000000    2.000000    1.000000
     2  n_changed      13.000000   20.000000   33.000000   20.000000   10.000000    2.000000    1.000000
     3  n_changed      11.000000   20.000000   31.000000   20.000000   10.000000    0.000000    1.000000
     4  n_changed      12.000000   20.000000   32.000000   20.000000    9.000000    2.000000    1.000000
     5  n_changed      13.000000   19.000000   32.000000   19.000000   10.000000    2.000000    1.000000
```

Each input area (sample) is represented by a row, whilst each cell type is represented by a column. Cell 'type' = no. of cells on which a cell's value depends. (Please note that the column headed cell type 1 is the direct equivalent of the column headed 'internal'.)

If two measures of tabular fit are requested (no. and % of table cells changed by disclosure control), the output will look like:

```
=== Table-specific, Area-specific, Table-based report for s06a ===

            Cell type (no. of contributing cells count depends upon)
Sample  Measure       Marginal     Internal    All         1          2          10         20
     1  n_changed      14.000000   17.000000   31.000000   17.000000   11.000000    2.000000    1.000000
     1  p_changed     100.000000   77.272727   86.111111   77.272727  100.000000  100.000000  100.000000
     2  n_changed      13.000000   20.000000   33.000000   20.000000   10.000000    2.000000    1.000000
     2  p_changed      92.857143   90.909091   91.666667   90.909091   90.909091  100.000000  100.000000
     3  n_changed      11.000000   20.000000   31.000000   20.000000   10.000000    0.000000    1.000000
     3  p_changed      78.571429   90.909091   86.111111   90.909091   90.909091    0.000000  100.000000
     4  n_changed      12.000000   20.000000   32.000000   20.000000    9.000000    2.000000    1.000000
     4  p_changed      85.714286   90.909091   88.888889   90.909091   81.818182  100.000000  100.000000
     5  n_changed      13.000000   19.000000   32.000000   19.000000   10.000000    2.000000    1.000000
     5  p_changed      92.857143   86.363636   88.888889   86.363636   90.909091  100.000000  100.000000
```

and so on.

5b(iv)Table-specific, Cross-area, Cell-based: summarises the distribution of user-requested cell-based measures across all input areas (samples), on a table-by-table basis. For example, the user might require the mean and maximum percentage change in a cell-based value across all user-supplied input areas arising from disclosure control:

```
=== Table-specific, Cross-area, Cell-based report (user-requested); s71 ===

original_cnt Maximum 10426.00000   152.00000 10191.00000   411.00000  3789.00000   224.00000
original_cnt Mean     9746.90000    54.20000  9383.60000   309.10000  3682.40000   147.60000

original_cnt Maximum   997.00000    14.00000   997.00000     0.00000   380.00000     0.00000
original_cnt Mean      931.70000     5.20000   926.50000     0.00000   368.70000     0.00000


cell_changed Maximum     1.00000     1.00000     1.00000     1.00000     1.00000     1.00000
cell_changed Mean        0.80000     0.90000     0.80000     1.00000     0.90000     0.90000

cell_changed Maximum     1.00000     1.00000     1.00000     0.00000     1.00000     0.00000
cell_changed Mean        1.00000     0.60000     1.00000     0.00000     0.90000     0.00000
```

As for *table-specific, area-specific, cell-based* reports (5b(ii)), the layout of cells conforms to the layout of cells in the user-supplied input tables (in this case, a table comprising one row and six columns).

The full range of cellular measures and distributional summary statistics available are set out below (see section 5(c) below headed 'Cell-based measures').

If multiple distributional measures are requested, including the mean, the report output will include report the mean twice: once in conjunction with the other requested measures, as illustrated above, and once in a stand-alone section, as illustrated below:

```
=== Table-specific, Cross-area, Cell-based report (mean); s71 ===

original_cnt Mean     9746.90000    54.20000  9383.60000   309.10000  3682.40000   147.60000
cell_changed Mean        0.80000     0.90000     0.80000     1.00000     0.90000     0.90000

original_cnt Mean      931.70000     5.20000   926.50000     0.00000   368.70000     0.00000
cell_changed Mean        1.00000     0.60000     1.00000     0.00000     0.90000     0.00000
```

If produced, the stand-alone 'mean' section precedes the section containing all requested distributional measures. This feature is designed to aid summary results analysis.

5b(v) Table-specific, Cross-area, Table-based: summarises the distribution of user-requested table-based measures across all input areas (samples), on a table-by-table basis. For example, the user might require the mean, maximum and minimum, across all user-supplied input areas, of the number and percentage of cells changed within each user-supplied input table as a result of disclosure control:

```
=== Table-specific, Cross-area, Table-based report (user-requested); s71 ===

                    Cell type (no. of contributing cells)
Measure    Distrib    Marginal    Internal         All           1           3
n_changed  Maximum    2.000000    8.000000   10.000000    8.000000    2.000000
n_changed  Mean       1.800000    7.000000    8.800000    7.000000    1.800000
n_changed  Minimum    1.000000    6.000000    8.000000    6.000000    1.000000
p_changed  Maximum  100.000000   80.000000   83.333333   80.000000  100.000000
p_changed  Mean      90.000000   70.000000   73.333333   70.000000   90.000000
p_changed  Minimum   50.000000   60.000000   66.666667   60.000000   50.000000
```

Note that, as for *table-specific, area-specific, table-based* reports (see 5b(iii) above), each table is considered as comprising a number of 'versions', each based on aggregations of cells of the same 'type'. A separate column is produced for each table cell type.

The full range of tabular measures and distributional summary statistics available are set out below (see section 5(d) below headed 'Table-based measures').

If multiple distributional measures are requested, including the mean, the report output will include report the mean twice: once in conjunction with the other requested measures, as illustrated above, and once in a stand-alone section, as illustrated below:

```
=== Table-specific, Cross-area, Table-based report (mean); s71 ===

                    Cell type (no. of contributing cells)
Measure    Distrib    Marginal    Internal         All           1           3
n_changed  Mean       1.800000    7.000000    8.800000    7.000000    1.800000

p_changed  Mean      90.000000   70.000000   73.333333   70.000000   90.000000
```

Note that distributional information is not available for the optional tabular measure 'frequency', which provides a simple count of the number of cells of each type in a table. Consequently, if this measure is requested, it will effectively be added as an additional header row. For example:

```
=== Table-specific, Cross-area, Table-based report (user-requested); s71 ===

                    Cell type (no. of contributing cells)
Measure    Distrib    Marginal    Internal         All           1           3
frequency  Count             2          10          12          10           2
n_changed  Maximum    2.000000    8.000000   10.000000    8.000000    2.000000
n_changed  Mean       1.800000    7.000000    8.800000    7.000000    1.800000
n_changed  Minimum    1.000000    6.000000    8.000000    6.000000    1.000000
p_changed  Maximum  100.000000   80.000000   83.333333   80.000000  100.000000
p_changed  Mean      90.000000   70.000000   73.333333   70.000000   90.000000
p_changed  Minimum   50.000000   60.000000   66.666667   60.000000   50.000000
```

5b(vi) Area-specific, Cross-table, Table-based: a report of user-specified table-based measures, averaged across all user-supplied input tables. The report layout follows that of area-specific, table-specific, table-based reports, with measures calculated separately for each cell type. Hence, tabular measures reported for in the column headed '4' represent the cross-table average of all marginal cells dependent upon the values of four internal cells. The results are reported separately for each user-supplied input area (sample):

```
=== Cross-table, Area-specific, Table-based report ===

                  Cell type (no. of contributing cells)
   Measure  Sample    Marginal    Internal         All           1           2           3          10          20
   n_changed     1   16.000000   25.000000   41.000000   25.000000   11.000000    2.000000    2.000000    1.000000
   n_changed     2   15.000000   26.000000   41.000000   26.000000   10.000000    2.000000    2.000000    1.000000
   n_changed     3   13.000000   28.000000   41.000000   28.000000   10.000000    2.000000    0.000000    1.000000
   n_changed     4   14.000000   28.000000   42.000000   28.000000    9.000000    2.000000    2.000000    1.000000


                  Cell type (no. of contributing cells)
   Measure  Sample    Marginal    Internal         All           1           2           3          10          20
   p_changed     1  100.000000   78.125000   85.416667   78.125000  100.000000  100.000000  100.000000  100.000000
   p_changed     2   93.750000   81.250000   85.416667   81.250000   90.909091  100.000000  100.000000  100.000000
   p_changed     3   81.250000   87.500000   85.416667   87.500000   90.909091  100.000000    0.000000  100.000000
   p_changed     4   87.500000   87.500000   87.500000   87.500000   81.818182  100.000000  100.000000  100.000000
```

5b(vii) Cross-table, Cross-area, Table-based: this report summarises user-specified measures of tabular fit across all user-supplied input areas (samples) and all user-supplied input tables. Summary and tabular measures reported are specified by the user. A full list of the tabular and summary measures available is listed below (5d(i)). The report output format follows that of *table-specific, area-specific, table-based* reports (5b(iii)), with a separate output column for each table cell type.

For example:

```
=== Cross-table, Cross-area, Table-based report (user requested) ===

                  Cell type (no. of contributing cells)
Measure    Distrib    Marginal    Internal         All           1           2           3          10          20
frequency  Count            16          32          48          32          11           2           2           1
n_changed  Maximum    16.000000   28.000000   42.000000   28.000000   11.000000    2.000000    2.000000    1.000000
n_changed  Mean       14.400000   26.000000   40.400000   26.000000   10.000000    1.800000    1.600000    1.000000
n_changed  Minimum    13.000000   24.000000   38.000000   24.000000    9.000000    1.000000    0.000000    1.000000
p_changed  Maximum   100.000000   87.500000   87.500000   87.500000  100.000000  100.000000  100.000000  100.000000
p_changed  Mean       90.000000   81.250000   84.166667   81.250000   90.909091   90.000000   80.000000  100.000000
p_changed  Minimum    81.250000   75.000000   79.166667   75.000000   81.818182   50.000000    0.000000  100.000000
```

reports the mean, maximum and minimum, across all user-supplied areas and tables, of the number and percentage of table cells changed by disclosure control.

If multiple distributional measures are requested, including the mean, the report output will include report the mean twice: once in conjunction with the other requested measures, as illustrated above, and once in a stand-alone section, as illustrated below:

```
=== Cross-table, Cross-area, Table-based report (mean) ===

                  Cell type (no. of contributing cells)
Measure    Distrib    Marginal    Internal         All           1           2           3          10          20
frequency  Count            16          32          48          32          11           2           2           1
n_changed  Mean       14.400000   26.000000   40.400000   26.000000   10.000000    1.800000    1.600000    1.000000
p_changed  Mean       90.000000   81.250000   84.166667   81.250000   90.909091   90.000000   80.000000  100.000000
```

5b(viii) Correct Rank: If this flag is switched on, and *use counts/percentages* <> 0, a report is generated indicating the extent to which the ranking of input areas by observed (post-disclosure control) percentages matches the ranking of input areas by expected (original) percentages. The process of ranking and assessment of correct rank is repeated for each percentage identified via percentage mapping (see (8) below).

An example of the output produced, for two percentages only, follows. Subsequent percentages would appear as additional columns in the output. To aid readability, the example output below has been edited to ensure column alignment. The raw space-separated output is best viewed, particularly when many percentages are involved, via a spreadsheet.

```
=== Correct Rank; percentages ===

pltill      pltill  pltill    punemp      punemp  punemp
CorrectRank Samples %_correct CorrectRank Samples %_correct
6           10      60.00     10          10      100.00
```

In *SDC_Direct_Impacts*, 'Samples' is synonymous with input areas. Hence the above output shows that, when ranked by % illness (pltill), 6 out of 10 areas (60%) had the same ranking pre- and post-disclosure control.

The report *Correct Rank* appears between any table-specific and cross-table reports requested.

N.B. In the case of areas with identical values, all are assigned the rank of the first occurring instance of the value, with the next occurring value having a rank = to this rank + no. of duplicate values. Ranking is from lowest to highest value, with rank 1 equalling lowest value.

E.g. Values in ascending order	Assigned rank
0.1	1
0.2	2
0.4	3
0.4	3
0.5	5

5b(ix) Correct Class: If this flag is switched on, and *use counts/percentages* <> 0, the number of areas placed into the same pre- and post-disclosure control quantiles (classes) is reported, for each of three quantile types: 20/10/5. For each quantile type the report commences by identifying the relevant upper and lower class boundaries. This is followed by an assessment of classification by individual class, which is followed in turn by an overall assessment.

Example output is given below for only two percentages – additional percentages would appear in additional columns. Edited here to ensure column alignment, this space-separated output is best viewed by via a spreadsheet.

```
=== Quantile boundaries ( 5  classes); percentages ===

Percentile:   20  class:  1  Lower-bound:  1  Upper-bound:   2
Percentile:   40  class:  2  Lower-bound:  3  Upper-bound:   4
Percentile:   60  class:  3  Lower-bound:  5  Upper-bound:   6
Percentile:   80  class:  4  Lower-bound:  7  Upper-bound:   8
Percentile:  100  class:  5  Lower-bound:  9  Upper-bound:  10

=== Correct Class ( 5  quantiles); percentages ===

Percentage pltill        pltill        pltill       punemp        punemp        punemp
Class      Correct_Class no._in_class  %_Correct    Correct_Class no._in_class  %_Correct
1          1              2             50.00        2             2             100.00
2          0              2              0.00        2             2             100.00
3          1              2             50.00        2             2             100.00
4          2              2            100.00        2             2             100.00
5          2              2            100.00        2             2             100.00

All        Correct_Class no._in_sample %_Correct    Correct_Class no._in_sample %_Correct
classes    6             10             60.00       10            10            100.00

=== Quantile boundaries ( 10  classes); percentages ===

Percentile:   10  class:   1  Lower-bound:   1  Upper-bound:   1
Percentile:   20  class:   2  Lower-bound:   2  Upper-bound:   2
Percentile:   30  class:   3  Lower-bound:   3  Upper-bound:   3
Percentile:   40  class:   4  Lower-bound:   4  Upper-bound:   4
Percentile:   50  class:   5  Lower-bound:   5  Upper-bound:   5
Percentile:   60  class:   6  Lower-bound:   6  Upper-bound:   6
Percentile:   70  class:   7  Lower-bound:   7  Upper-bound:   7
Percentile:   80  class:   8  Lower-bound:   8  Upper-bound:   8
Percentile:   90  class:   9  Lower-bound:   9  Upper-bound:   9
Percentile:  100  class:  10  Lower-bound:  10  Upper-bound:  10

Etc…
```

The report *Correct Class* appears between any table-specific and cross-table reports requested.

5b(x) Correct/Neighbouring Class: If this flag is switched on, and *use counts/percentages* <> 0, the number of areas placed into the same or an adjacent pre- and post-disclosure control quantile (class) is reported, for each of three quantile types: 20/10/5. For each quantile type the report commences by identifying the relevant upper and lower class boundaries. This is followed by an assessment of classification by individual class, which is followed in turn by an overall assessment.

Example output is given below for only two percentage – additional percentages would appear in additional columns. Edited here to ensure column alignment, this space-separated output is best viewed by via a spreadsheet. The column headed 'Near_Class' records the number of observed input areas falling within the relevant, or an adjacent, class.

```
=== Quantile boundaries ( 5  classes); percentages ===

Percentile:   20  class:  1  Lower-bound:  1  Upper-bound:   2
Percentile:   40  class:  2  Lower-bound:  3  Upper-bound:   4
Percentile:   60  class:  3  Lower-bound:  5  Upper-bound:   6
Percentile:   80  class:  4  Lower-bound:  7  Upper-bound:   8
Percentile:  100  class:  5  Lower-bound:  9  Upper-bound:  10

=== Correct/Neighbouring class ( 5  quantiles); percentages ===

Percentage pltill     pltill        pltill      punemp     punemp        punemp
Class      Near_Class no._in_class  %_Correct   Near_Class no._in_class  %_Correct
1          2           2            100.00       2          2            100.00
2          2           2            100.00       2          2            100.00
3          2           2            100.00       2          2            100.00
4          2           2            100.00       2          2            100.00
5          2           2            100.00       2          2            100.00

=== Quantile boundaries ( 10  classes); percentages ===

All        Near_Class no._in_sample %_Correct   Near_Class no._in_sample %_Correct
classes    10         10            100.00      10         10            100.00

Percentile:   10  class:   1  Lower-bound:   1  Upper-bound:   1
Percentile:   20  class:   2  Lower-bound:   2  Upper-bound:   2
Percentile:   30  class:   3  Lower-bound:   3  Upper-bound:   3
Percentile:   40  class:   4  Lower-bound:   4  Upper-bound:   4
Percentile:   50  class:   5  Lower-bound:   5  Upper-bound:   5
Percentile:   60  class:   6  Lower-bound:   6  Upper-bound:   6
Percentile:   70  class:   7  Lower-bound:   7  Upper-bound:   7
Percentile:   80  class:   8  Lower-bound:   8  Upper-bound:   8
Percentile:   90  class:   9  Lower-bound:   9  Upper-bound:   9
Percentile:  100  class:  10  Lower-bound:  10  Upper-bound:  10

Etc…
```

The report *Correct/Neighbouring Class* appears between any table-specific and cross-table reports requested.

5(c) Cell-based measures

[For each measure of fit, 0='off'; 1='on']

5c(i) Measures available

SDC_Impact_Direct calculates, and can report if required, 8 cell-based measures. (Note that to report cell-based measures a cell-based report-type must also have been requested.)

cell_exp: expected cell value (original value)

cell_obs: observed cell value (value after application of disclosure control)

cell_changed: A flag indicating whether expected and observed cell values differ (1=differ; 0=no difference)

cell_TE: Total Error (size of difference between expected and observed values)

cell_Z: Z-score (depends upon size of difference and table total; see p.38 for details)

cell_NFC: Flag set to '1' if cell | Z-score | is > 1.96, indicating a 'non-fitting cell' [i.e. difference between expected and observed count greater than would be expected by change alone (0.05 significance level)]; else flag set to '0'.

cell_Zm: Modified Z-score (Z_m)which takes account of cases when expected and observed table totals are markedly different (see appendix p.38 for details).

cell_NFCm: Flag set to '1' if cell | Z_m | is > 1.96, indicating a 'non-fitting cell'; else flag set to '0'. [modified Z does not have a known sampling distribution, although if expected table total = observed table total, $Z_m = Z$]

5c(ii) Cross-area summary values available

For each cell-based measure, five sample summary values are available:

Cell_Summary, Max: Maximum value of cell-based measure across all input areas

Cell_Summary, 97.5%-tile: 97.5th percentile-value of cell-based measure across all input areas

Cell_Summary, mean: mean value of cell-based measure across all input areas

Cell_Summary, 2.5%-tile: 2.5th percentile-value of cell-based measure across all input areas

Cell_Summary,min: Minimum value of cell-based measure across all input areas

5(d) Table-based measures

In (i) and (ii) below the term 'table' is used in the sense outlined in more detail in section (iii). Full definitions of all measures are given in pages 38-41. The measures listed below will only be reported if a 'table-based' report type has also been requested.

5d(i) Available measures of tabular fit

SDC_Direct_Impact produces the following range of measures of tabular fit:

Table_frequency (of cell type): No. of cells in a table of a given 'type' [see (iii) below]

Table_n_changed: No. of cells in table who's expected (original) and observed (post disclosure control) values differ

Table_p_changed: % of cells in table who's expected and observed values differ

Table_max_change: Maximum difference (change) in pre- and post-disclosure control cell values

Table_maxPchange: Maximum % difference (change) in pre- and post-disclosure control cell values

Table_TotalError: Total Error - difference between expected and observed counts summed across all table cells

Table_TAE: Total Absolute Error - absolute difference between expected and observed counts summed across all table cells

Table_RAE: Relative Absolute Error – TAE as % of total value of changed cells

Table_SAE: Standardised Absolute Error – TAE / sum of table cells (table total)

Table_Sq_Error: Total Square Error – sum of square of difference between expected and observed cell values

Table_RMSE: Square root of the average square error across all table cells.

Table_SSZ: Sum of the square of the cell Z-scores

Table_NFC: No. of 'Non-Fitting Cells' in table. [i.e. no. of cells with | Z-score | > 1.96] (i.e. no. of cells for which difference between expected and observed values is greater than can be explained by chance at the 0.05 significance level).

Table_NFT: Non-fitting table; ='1' if table SSZ exceeds critical value (at 0.05 significance level); else = 0

Table_SSZm: Sum of the square of the cell modified Z-scores [see p.38 A for full explanation of Z_m)

Table_NFCm: No. of 'Non-Fitting Cells' in table [i.e. no. of cells with | Z_m-score | > 1.96] (N.B. value of 1.96 is arbitrary as Zm has no known sampling distribution unless expected and observed table totals are the same).

Table_NFTm: Non-fitting table; ='1' if table SSZ_m exceeds SSZ critical value (at 0.05 significance level); else = 0 (SSZ_m has unknown sampling distribution unless expected and observed table totals are the same)

Table_Gibsons_D: Gibson's D

Table_Cramers_V: Cramer's V

Table_PearsonsR: Pearsons Correlation Coefficient

Table_ChiSquare: Chi-square

Table_TVCC: Total expected value of all cells for whom expected and observed values differ

Table_v_expcells: Sum of expected cell values

Table_v_obscells: Sum of observed cell values

5d(ii) Cross-area five sample summary values are available

Table_Summary, Max: Maximum value of table-based measure across all input areas
Table_Summary, 97.5%-tile: 97.5th percentile-value of table-based measure across all input areas
Table_Summary, mean: mean value of table-based measure across all input areas
Table_Summary, 2.5%-tile: 2.5th percentile-value of table-based measure across all input areas
Table_Summary, min: Minimum value of table-based measure across all input areas

(e) 'Tables' and 'cell types'

Conventionally, measures of tabular fit are based on a table's internal cells (i.e. all cells whose value depends on no other cell). However, in terms of disclosure control, the cumulative impact on marginals is of particular interest. For this reason, *SDC_Direct_Impact* produces 'table-based' measures based on evaluation not only of all internal cells, but also, separately, for all cells of a given 'type' within each table. A cell's 'type' is defined by the number of other cells within the table upon which it's value depends. Internal cells are type '0' (their values depend on no other cells). In contrast, cells of type 4 represent all marginal cells in a table whose value depends upon the summation of 4 internal cells. In addition, two other cell types are also recognised: all cells, whether marginal or internal, denoted by cell type '-2'; and all marginal cells (i.e. all cells depending on the value of 1+ other cells), denoted by cell type '-1'. During calculation a 'table' is regarded as comprising all table cells of a given 'type'. Please note that, for internal programming reasons, all cells reported in all *SDC_Direct_Impacts* output as cells of type 1 are, in fact, cells of type 0 [i.e. type 1 = internal cells]. This is because cells of type 1, depending on only 1 cell are, in effect, simply direct copies of existing internal (type 0) cells.

6) *SDC_Direct_Impacts_Count_input_tables.fmt*

[Stored in the *RunParameters* folder pointed to in *Input_and_output_paths.txt*]

A list of files containing lists of pre/post perturbation table counts to be used in assessment of disclosure control (one pair of comparison tables per row of file).

The format for each comparison pair (row) in the file is:
"*<table name>*", *<original count variant>*, *<perturbed count variant>*

E.g.

```
"s06", 0, 2
```

It is important that: (i) the table name is in quotes; (ii) all items in the row are comma-separated; (iii) the *table name* supplied matches the *table name* used in the naming of input and map files (see (1), (2) and (3) above if in doubt).

The file *SDC_Disclosure_Impacts_run_parameters.txt* contains all additional information required to generate full input file names covering both map files and original/perturbed count data, regardless of data source (user-supplied, or created via *Create_Aggregates*).

For a user-supplied set of tables, the example given above is equivalent to requesting that the counts contained in the file

*S06_v**0**.fmt*

are compared to their equivalents in

*S06_v**2**.fmt*

If the data source for the tables is *Create_Aggregates*, the example above is equivalent to requesting that the counts contained in the file

*S06a_v**0**_P20[Popdens]_n20[R]_s1000.fmt*

are compared to their equivalents in

*S06a_v**2**_P20[Popdens]_n20[R]_s1000.fmt*

7) *SDC_Direct_Impacts_Percentage_input_tables.fmt*

[Stored in the *RunParamters* folder pointed to in *Input_and_output_paths.txt*]

If the *Use counts/percentages* option has been set to 1 or 2 in *SDC_Direct_Impacts_run_parameters.txt*, then this file is required as input. The file should list files containing pre/post perturbation table *percentages* to be used in assessment of disclosure control (one pair of comparison tables per row of file). For example,

The format for each comparison pair (row) in the file is:
"*<table name>*", *<original count variant>*, *<perturbed count variant>*

E.g.

```
"percentages", 0, 2
```

It is important that: (i) the table name is in quotes; (ii) all items in the row are comma-separated; (iii) the *table name* supplied matches the *table name* used in the naming of input and map files (see (1), (2) and (3) above if in doubt).

SDC_Direct_Impacts will parse root table name(s) into full input filename(s) in precisely the same manner as for files containing count data, as outlined for *SDC_Direct_Impacts_Count_input_tables.fmt* above.

8) *<percentage name>.map*

[Located in the *TableMappings* folder]

If the *Use counts/percentages* option has been set to 1 or 2 in *SDC_Direct_Impacts_run_parameters.txt*, then this file is required as input (one map file per input file listed in *SDC_Direct_Impacts_Percentage_input_tables.txt*).

This file describes the format of the associated percentage input file. Just as for count data, percentage data can be supplied in tabular or vector format. The first line of the file <percentage name>.map describes the number of rows and columns per input area.

For example

```
1 17
```

describes an input file with 17 percentages per input area, laid out as a vector (1 row).

For percentages whose value depends on the summation of other percentages, additional mapping information is required, just as for count data (see section 3 'Table Mappings' above).

9) *Chisquare.dat*

[Stored in the *RunParamters* folder pointed to in *Input_and_output_paths.txt*]

A file, supplied with the program, that gives chi-square critical values, at 0.05 significance level, for 0 to 5000 degrees of freedom. Needed to check whether or not pre- and post-disclosure counts agree at the tabular level, using squared Z-score (which has unit normal distribution).

PROGRAM OUTPUTS

SDC_Direct_Impacts_results.txt

[Stored in the folder pointed to by *ProgramPath*]

All output from *SDC_Direct_Impacts* is written to this file. The precise contents of the output depend upon the reports requested by the user via *SDC_Direct_Impacts_run_parameters.txt*. Details of the output produced by each report are given under the relevant report heading in section 5 of *Program Inputs* above. More complex output may best be viewed via a spreadsheet package. For the purpose of importing to a spreadsheet package, the program output should be regarded as space-separated.

FULL DESCRIPTIONS OF TABULAR AND CELLULAR MEASURES

(1) Cellular measures for count data

Definitions

Cell type - the number of internal cell counts on which a cell's value is based. Internal cells have a cell type of 0; marginal cells have a cell type of 2 or more. Cells of type 1 are direct copies of internal cells, and are treated as internal cells for classification purposes.

Cell [i] = specific cell within table (i ranges from 1 to number of cells in table)

Measures

Exp [E_i] = expected (pre disclosure control) cell value

Obs [O_i] = observed (post disclosure control) cell value (value after application of disclosure control)

Changed [C_i] = 1 if $O_i <> E_i$; else = 0.

TE [TE_i] = $O_i - E_i$

Z [Z_i] = $[(O_i / \Sigma O_i) - (E_i / \Sigma E_i) + Q_i] / [\{ (E_i / \Sigma E_i)(1-(E_i / \Sigma E_i)) \} / \Sigma O_i]^{0.5}$,

where $Q_i = 0$ if $E_i = 0$; else if $(O_i / \Sigma O_i) - (E_i / \Sigma E_i) > 0$, $Q_i = -(1/(\Sigma E_i + \Sigma O_i))$;
else $Q_i = +(1/(\Sigma E_i + \Sigma O_i))$.

To avoid Z_i becoming undefined:

(i) if $E_i = 0$, substitute $E_i = 1$
(ii) if $E_i = \Sigma E_i$, substitute ΣE_i with $\Sigma E_i + 1$
(iii) if $Ei > \Sigma E_i$, substitute ΣE_i with $E_i + 1$
(iv) if $E_i = O_i$ and $\Sigma E_i = \Sigma O_i$, $Z_i = 0$

NFC [NFC_i] = 1 if $| Z_i |$ exceeds critical value of 1.96 (p=0.05); else 0.

Zm [Zm_i] = $[(O_i / \Sigma E_i) - (E_i / \Sigma E_i)] / [\{ (E_i / \Sigma E_i)(1-(E_i / \Sigma E_i)) \} / \Sigma E_i]^{0.5}$

To avoid Zm_i becoming undefined:

(i) if $E_i = 0$, substitute $E_i = 1$
(ii) if $E_i = \Sigma E_i$, substitute ΣE_i with $\Sigma E_i + 1$
(iii) if $Ei > \Sigma E_i$, substitute ΣE_i with $E_i + 1$
(iv) if $E_i = O_i$ and $\Sigma E_i = \Sigma O_i$, $Zm_i = 0$

NFCm [$NFCm_i$] = 1 if $| Zm_i | > 1.96$; else 0

(2) Tabular measures for count data

Definitions

Table – input tables will typically comprise a set of internal cell counts, possible plus a set of table margins. It is possible to envisage assessing the impact of disclosure control on all table cells, on internal cells only, on marginal cells only and so on. For analytical purposes, therefore, a 'table' is taken to represent a set of cells of common cell type (e.g. all marginal cells based on the summation of 4 internal cells). In consequence one input table may have generate multiple 'table' outputs.

Measures

frequency (n) = a count of the number of cells within a given table

n_changed (NC) = $\Sigma\, NC_i$, where $NC_i = 1$ if $O_i <> E_i$; 0 otherwise.

> O = observed (post-disclosure control) counts; E = expected (pre-disclosure control) counts; i = specific cell within table.

p_changed (PC) = $(\Sigma\, NC_i) / n$

max_change (MNC) = max $(O_i - E_i)$, for $i = 1$ to n

maxPchange (MPC) = max $\{(O_i - E_i)/E_i\}$, for $i = 1$ to n

TotalError [TE] = $\Sigma\, (O_i - E_i)$, for $i = 1$ to n

TAE (TAE) = $\Sigma\, |\, (O_i - E_i)\, |$, for $i = 1$ to n

RAE (RAE) = $100(TAE_i / TVC)$ [%] [see below for definition of TVC]

SAE [SAE] = $TAE / (\Sigma\, E_i)$, for $i = 1$ to n

Sq_Error [E^2] = $\Sigma\, (O_i - E_i)^2$, for $i = 1$ to n

RMSE [$RMSE$] = $(E^2 / n)^{0.5}$

SSZ [SSZ] = $\Sigma\, Z_i^2$, for $i = 1$ to n

NFC [NFC] = $\Sigma\, NFC_i$, *for i = 1 to n*

NFT [NFT] = 1 if SSZ exceeds χ^2 critical value for table (p=0.05; $df = n$); else 0.

(i) Degrees of freedom: calculation of NFT assumes that all cells, internal and marginal, are not constrained in their fit to pre-disclosure control values. Hence degrees of freedom, for any table, is taken to be n.

This stance is justified as follows. First, few, if any, disclosure control methods currently implemented by statistical agencies involve modifying internal cells in such a way that they are guaranteed to total to original marginals. Such a method would, in any case, probably open up the possibility of reverse-engineering the perturbations applied. Consequently, in assessing degrees of freedom, all internal cells may be regarded as unconstrained. If post

disclosure control marginal values are also not constrained, then the assumption that df = n remains valid. However, it is possible that margins are independently supplied and constrained to fit to original margins, in which case degrees of freedom for marginal cells = 0. If this is the case the values of NFT for all cell types except internal should be disregarded.

SSZm [$SSZm$] = $\Sigma\ Zm_i^2$, for i = 1 to n

NFCm [$NFCm$] = $\Sigma\ NFCm_i$, *for i = 1 to n*

NFTm [$NFTm$] = 1 if $SSZm$ exceeds χ^2 critical value for table (p=0.05; $df = n$); else 0.

Gibsons_D [D] = 0.5 Σ | $(E_i / \Sigma E_i) - (O_i / \Sigma O_i)$ | , for i = 1 to n

(i) If $\Sigma E_i = 0$, set $E_i / \Sigma E_i = 0$; if $\Sigma O_i = 0$, set $O_i / \Sigma O_i = 0$

Cramers_V [V] = [$\chi^2 / n \min(r\text{-}1, c\text{-}1)$] $^{0.5}$,

where r = no. of rows (of given cell type) in table; c = no. of columns in table (in table).

(i) If minimum (r, c) =1, V = -9 [undefined]

(ii) For cell types other than internal, the value of V represents only an approximate measure of fit

PearsonsR [r] = $\Sigma\ [(O_i - O_m)(E_i - E_m)] / [\ \Sigma(O_i - O_m)^2\ \Sigma(E_i - E_m)^2]^{0.5}$, for i = 1 to n,

where $O_m = \Sigma\ O_i / n$ and $E_m = \Sigma\ E_i / n$

(i) If $\Sigma(O_i - O_m)^2 = 0$ or $\Sigma(O_i - O_m)^2 = 0$, set $r = 0$

(ii) If number of cells in table = 1, r = -9 [undefined]

ChiSquare [χ^2] = $\Sigma\ \{\ (O_i - E_i)^2 / E_i\ \}$, for i = 1 to n

TVCC [$TVCC$] = $\Sigma\ E_i$, for all i where $E_i <> O_i$

v_expcells [ΣE_i] = $\Sigma\ E_i$, for i =1 to n

v_obscells [ΣO_i] = $\Sigma\ O_i$, for i =1 to n

(3) Cross-table measures for count data

In definitions given in this section, $\Sigma\ X$ = sum indicated measure (X) across all input tables

N_changed	$\Sigma\ NC$
P_changed	$\Sigma\ NC / \Sigma\ n$
Max_change	Maximum MNC
MaxPchange	Maximum MPC
TotalError	$\Sigma\ TE$
TAE	$\Sigma\ TAE$

RAE	100(Σ *TAE* / Σ *TVCC*)
SAE	Σ *TAE* / Σ *Ei*, for *i*= 1 to Σ*n*
SqError	Σ E^2
RMSE	Σ *RMSE*
SSZ	Σ *SSZ*
NFC	Σ *NFC*
NFT	Σ *NFT*
SSZm	Σ *SSZm*
NFCm	Σ *NFCm*
NFTm	Σ *NFTm*
GibsonsD	As for tabular measure, but for *i* = 1 to Σ *n*
Cramers_V	*V* / *T*, where *T* = no. of tables [an approximation required because min (*r*-1, *c*-1) is a meaningless concept across multiple tables]
PearsonsR	As for tabular measure, but for *i* = 1 to Σ *n*
ChiSquare	Σ χ^2 [*df* = Σ *n*]
TVCC	Σ *TVCC*
v_expcells	As for tabular measure, but for *i* = 1 to Σ *n*
v_obscells	As for tabular measure, but for *i* = 1 to Σ *n*

(4) Measures for use with percentages

The following measures of fit are inappropriate for use with percentage data:

Cellular: *Z*, *NFC*, *Zm*, *NFCm*
Tabular: *SSZ*, *NFC*, *NFT*, *SSZm*, *NFCm*, *NFTm*, *V*, χ^2

Therefore, even if requested, *SDC_Direct_Impacts* will not report these measures for percentage data.

(5) Distributional measures

Available measures: Maximum, minimum, mean, 5th and 95th percentiles.

Percentiles – calculated by interpolation given Q, Q = rank of value for given percentile. Q = 1+(p(N-1)), where p = percentile required, expressed as a fraction) (e.g. 0.95 = 95th percentile) and N = no. of ranked values (i.e. no. of input areas).

SDC_Indirect_Impacts

[Unless stated, all file names in this section are given relative to *D:\Research\Disclosure Control\ Disclosure_VB\SDC_Indirect_Impacts_v9*]

Output from *Create_Aggregates* can include a set of area-specific percentages, produced from a set of user-supplied input data. The precise nature of the areas used for output by *Create_Aggregates* depends on a wide range of user-specified options, including whether or not the output areas are aggregates of one or more input areas, whether or not the output areas are randomly selected from strata within the input areas, and in what manner random samples of input areas are generated.

SDC_Indirect_Impacts takes the area-specific percentages (one row of percentages per input area) and:

- Identifies the first supplied percentage (first column) as the user-specified dependent variable
- Identifies all other supplied percentages (subsequent columns) as the user-specified indpendent variables
- Generates a number of random sub-sets (user-specified), each comprising a user- number of input areas (also user-specified). The same sub-set is taken from equivalent pre- and post-disclosure control input files.

For each sub-set, *SDC_Indirect_Impacts*:

- Runs a bivariate regression model for every combination of depedent and independent variable
- Runs a multiple regression using a hard-coded set of independent variables
- Identifies all regression and correlation signs that differ pre- and post-disclosure control
- Identifies all pre- and post-disclosure control regression coefficients that (i) differ by more than 1 standard error; (ii) exceed the relevant 95% confidence interval

Across all sub-sets, *SDC_Indirect_Impacts*:

- Assesses the % of regression and correlation signs that differ pre- and post-disclosure control (across all sub-sets of
- Assesses the % of pre- and post-disclosure control regression coefficients that (i) differ by more than 1 standard error; (ii) exceed the relevant 95% confidence interval
- Identifies the distribution (minimum, maximum, mean, 5^{th} and 95^{th} percentile) of the sub-set regression and correlation coefficient signs and differences

Note: Due to resource constraints (lack of time), this program currently has far less flexibility than SDC_Direct_Impacts, as many of program settings and limits are hard-coded. There is no technical reason why greater flexibility could not be introduced if the program code were appropriately revised.

Program limits

Percentages: 17
Input areas: 1000 (per input table variant)
Samples: 1000

Program inputs

1) *NoOfSample_sets.txt*

SDC_Indirect_Impacts reads in user-supplied tabular and/or percentage data, then randomly selects one or more sub-sets of these input areas. The subset(s) is/are used to assess the impact of disclosure control on a range of ecological analyses. This number of sub-sets generated are determined by the user-specified value given in this file.

E.g.

```
10
```

The number of areas in each sub-set is controlled by the separate file (see (2) below)

2) *SDC_Indirect_Impacts_sample_set_sizes.txt*

Number of user-supplied areas to be randomly selected by SDC*_Indirect_Impacts* when generating each sub-set of user-supplied areas.

E.g.

```
12
```

3) *Input file name(s)*

SDC_Indirect_Impacts.txt is currently only designed to work with output from *Create_Aggregates_v2*. In consequence, it expects files to be named following the convention:

*<tablename>***_v***<variant no.>*_*<strata type>***[***<strata source>***]_n***<sample size>***[***<sampling strategy>***]_s***<no. of samples>*.fmt**

E.g. *percentages_v3_P20[popdens]_n20[C]_s100.fmt*

SDC_Indirect_Impacts is capable of processing any number and combination of such input files, providing that, for each combination, a pre- [*v0*] and post- [*vX*] disclosure control version is available.

The following files supply the information required by *SDC_Indirect_Impacts* to identify its input files. Each row in a file represents one file type. Any number of rows, each representing a separate file type, are permissible, as *SDC_Indirect_Impacts* processes each file type in turn.

(a) *NoOfSamples.txt*

No. of input areas provided in each user-supplied file of input percentages.

E.g.

```
100
```

(b) *sampling_strategy.txt*

Way in which *Create_Aggregates* created (aggregate) samples of user-supplied areas [see *Create_Aggregates* documentation for details].

E.g.

```
c
```

(c) *sample_sizes.txt*

No. of areas in each sample of areas generated by *Create_Aggregates*.

E.g.

```
20
```

(d) *strata_names.txt*

Source of data used by *Create_Aggregates* for stratification of user-supplied input areas.

E.g.

```
popdens
```

(e) *strata_types.txt*

Range of strata types produced by Create_Aggregates

E.g.

```
P20
P80
```

(f) *sdc_variants.txt*

Input data post-disclosure control variants (assuming that v0 exists and represents pre-disclosure control variant).

E.g.

```
3
```

(g) *tablenames.txt*

Root name of table containing percentage data to be input to *SDC_Indirect_Impacts*.

E.g.

```
percentages
```

4) *SDC_Indirect_Impacts_run_parameters.txt*

The main output from *SDC_Indirect_Impacts* is written to the file *SDC_Indirect_Impacts_results.txt*. The contents of this output file are controlled by a number of user-specified report flags (1='on'; 0='off').

The range of flags available, and their impacts, are set out below. Many of the output options provide output best suited to checking that the end-results from SDC_Indirect_Impacts are being calculated correctly. The report option settings that produce the summary information of most potential interest to end-users are:

- *Report_summary_MR_signs_and_SEs*
- *Report_summary_R_signs_and_SEs*

Followed closely by

- *Report_MR_coeff_CIs*
- *Report_R_coeff_CIs*

(i) *report_input_pctgs*

If turned on, lists the names of all pre- and post-disclosure control input files read in, and echoes their contents. Columns = percentages; rows = input areas. This report type is principally designed for error checking only.

```
=== percentages_v3_P20[popdens]_n20[C]_s100.fmt ===

 Expected pctgs
pltill  punemp  pfultim pownocc plarent pnocent pusdst  pshared pnocarhh
12.591   8.321  35.040  68.629  26.346  20.600  90.689   0.139  34.483
13.594   5.617  34.231  76.897  16.010  20.296  84.414   0.123  29.648
.
.
.
  13.389   6.061  35.954  75.682  18.736  16.008  79.152   0.103  29.855  26.495   5.260   1.268
51.850   1.477   0.866   0.346   0.000

 Observed pctgs
pltill  punemp  pfultim pownocc plarent pnocent pusdst  pshared pnocarhh
  12.552   8.401  34.951  68.355  26.385  20.373  90.727   0.111  34.471
  13.506   5.484  34.143  76.768  15.840  20.408  84.481   0.123  29.425
  17.281  10.242  29.529  49.368  42.391  26.112  76.337   0.643  52.231
.
.
.
=== percentages_v3_P80[popdens]_n20[C]_s100.fmt ===

 Expected pctgs
pltill  punemp  pfultim pownocc plarent pnocent pusdst  pshared pnocarhh
  17.281  13.367  29.030  45.951  41.524  49.986  78.035   0.802  59.871
  15.795  11.508  32.219  59.351  25.495  58.141  90.080   0.249  57.210
.
.
.
etc.
```

(ii) *report_sample_set*

SDC_Indirect_Impacts generates a user-requested number of samples (sub-sets of areas) from the set of input areas supplied. To ensure direct comparability, the same sub-sets of areas are taken from a post-disclosure control input file and its pre-disclosure control equivalent. However, sub-set membership is not held constant across input files for differing strata (as an area can only belong to one strata); nor across input files generated generated using alternative statistical disclosure control methods.

If the *report_sample_set* option is turned on, the input areas comprising each randomly selected sub-set are listed, once for each percentage processed, and once prior to the processing of a multiple regression model. In the output produced, the second column gives the input area number (areas numbered in order of input); the first column gives sub-set membership no..

E.g.

```
=== percentages_v3_P20[popdens]_n20[C]_s100.fmt ===

 === punemp      Sample_set no. :     1

            1          19
            2          13
            3          40
            4           7
            5          68
            6          25
            7          19
            8          89
            9          73
           10          94
           11          59
           12          27

 === punemp      Sample_set no. :     2

            1          50
            2          94
            3          96
            4          32
            5          17
            6          53
            7          66
            8          24
            9          40
           10          67
           11          27
           12          81
.
.
.

 === pfultim     Sample_set no. :     1

            1          19
            2          13
            3          40
            4           7
            5          68
            6          25
            7          19
            8          89
            9          73
           10          94
           11          59
           12          27

 === pfultim     Sample_set no. :     2

            1          50
            2          94
            3          96
            4          32
            5          17
            6          53
            7          66
            8          24
            9          40
           10          67
           11          27
           12          81
.
.
.
```

```
===  Mult. Reg. Sample_set no. :     1

            1          19
            2          13
            3          40
            4           7
            5          68
            6          25
            7          19
            8          89
            9          73
           10          94
           11          59
           12          27
```

(iii) *report_MR_inputs*

If this option is switched on, the expected (pre-disclosure control) and observed (post-disclosure control) percentages are listed for each area selected as part of a sample (sub-set of all input areas). In the output, each column represents an input percentage (listed in the same order as input); each row represents an input area selected for sub-set membership. After each set of observed/expected percentages, a row of flags indicate which of the percentages have been selected for input to the regression model currently being processed [1='on'; 0='off']. The selection of independent variables used in the multiple regression model is currently hard-code into the program. This output is principally designed for error checking purposes.

```
=== percentages_v3_P20[popdens]_n20[C]_s100.fmt ===

 === punemp      Sample_set no. :     1

 Expected

 Input Y and X values
  12.874  12.874   5.380  32.976  76.181  16.735  21.612  78.771   0.413  30.720
  13.530  13.530   7.173  32.925  73.291  19.314  23.990  85.926   0.153  32.175
  13.594  13.594   5.617  34.231  76.897  16.010  20.296  84.414   0.123  29.648
  12.274  12.274   4.119  31.239  76.148  12.424  20.683  96.622   0.071  20.502
  16.455  16.455   9.330  31.463  55.329  36.295  28.186  77.144   0.741  48.526
  16.267  16.267   7.974  33.711  63.845  30.040  31.322  81.031   0.240  41.112
  12.874  12.874   5.380  32.976  76.181  16.735  21.612  78.771   0.413  30.720
  16.251  16.251   7.514  33.101  64.631  28.661  24.756  79.069   0.075  41.077
  15.804  15.804   8.344  31.499  57.487  32.227  32.379  74.987   0.279  45.305
  11.700  11.700   4.795  36.461  83.780  12.128  13.691  87.438   0.074  23.376
  15.945  15.945   7.649  31.784  63.119  29.283  19.436  79.482   0.085  38.597
  14.681  14.681   8.336  33.288  68.156  23.358  31.923  81.285   0.423  40.493

 Indep. Var. switches [1=on;0=off]
       0       1       0       0       0       0       0       0       0       0

 Observed

 Input Y and X values
  13.001  13.001   5.342  33.049  76.046  16.842  21.746  78.573   0.465  30.837
  13.542  13.542   7.232  33.031  73.452  19.236  23.745  85.908   0.102  32.344
  13.506  13.506   5.484  34.143  76.768  15.840  20.408  84.481   0.123  29.425
  12.369  12.369   4.177  31.133  76.441  12.384  20.641  96.825   0.000  20.355
  16.414  16.414   9.382  31.623  55.324  36.284  28.353  77.235   0.691  48.552
  16.230  16.230   7.994  33.752  63.801  30.032  31.527  80.922   0.213  41.170
  13.001  13.001   5.342  33.049  76.046  16.842  21.746  78.573   0.465  30.837
  16.254  16.254   7.535  33.042  64.794  28.761  24.624  79.243   0.075  40.997
  15.827  15.827   8.378  31.469  57.624  32.168  32.270  74.930   0.279  45.126
  11.739  11.739   4.888  36.535  83.988  12.108  13.749  87.382   0.124  23.473
  15.877  15.877   7.682  31.895  63.255  29.361  19.413  79.612   0.086  38.263
  14.600  14.600   8.363  33.342  67.965  23.348  32.140  81.308   0.397  40.616

 Indep. Var. switches [1=on;0=off]
       0       1       0       0       0       0       0       0       0       0
```

(iv) *report_detailed_MR_outputs*

SDC_Indirect_Impacts takes each sample (area sub-set) in turn and fits a series of bivariate regression models, in which the dependent variable is always the first percentage in the user-supplied percentage data (first column), and the independent variable is, in turn, each remaining user-supplied percentage. Finally, *SDC_Indirect_Impacts* fits a multiple-regression model, the independent variables of which are assigned via ****.

If this report option is switched on, detailed output is reported for each regression model fitted. E.g.

```
=== percentages_v3_P20[popdens]_n20[C]_s100.fmt ===

 === punemp      Sample_set no. :     1

 Expected

 Regression/correlation model outputs

 Resid sum of squares, RSS =           6.44012687299
 Degrees of freedom =           10

 Variable,   Parameter estimate  Standard eror

 Constant            0.7962E+01          0.1015E+01
 punemp              0.9400E+00          0.1452E+00

 R-square =  80.7274

 R        =   0.8985

 Observed

 Regression/correlation model outputs

 Resid sum of squares, RSS =           6.62506490814
 Degrees of freedom =           10

 Variable,   Parameter estimate  Standard eror

 Constant            0.8232E+01          0.1023E+01
 punemp              0.8994E+00          0.1461E+00

 R-square =  79.1293

 R        =   0.8895

.
.
.

===  Mult. Reg. Sample_set no. :     1

 Expected

 Regression/correlation model outputs

 Resid sum of squares, RSS =           4.43929871741
 Degrees of freedom =            6

 Variable,   Parameter estimate  Standard eror

 Constant           -0.1097E+02          0.1465E+02
 punemp              0.4769E+00          0.6608E+00
 pnocent             0.1108E-02          0.7108E-01
 pusdst              0.2249E+00          0.1953E+00
 p2carhh            -0.2393E+00          0.2311E+00
 page7584            0.1659E+01          0.1049E+01

 R-square =  86.7151

 R        =   0.0163

 Observed

 Regression/correlation model outputs
```

```
Resid sum of squares, RSS =           4.56756315049
Degrees of freedom =            6

Variable,   Parameter estimate  Standard eror

Constant           -0.6937E+01          0.1333E+02
punemp              0.5098E+00          0.6017E+00
pnocent             0.1306E-01          0.7069E-01
pusdst              0.1679E+00          0.1751E+00
p2carhh            -0.1769E+00          0.2038E+00
page7584            0.1416E+01          0.9242E+00

R-square =  85.6110

R        =   0.0163
```

(v) *report_summary_MR_outputs*

A summarised version of the regression model output set out in full by *report_detailed_MR_outputs* (above) is produced by this reporting option.

For each regression model, the following information is given, separated into four blocks.

The first blocks deals with results for regression models fitted to expected (pre-disclosure control) data; the second with regression models fitted to observed (post-disclosure control) data. Each block has the same format:

Model specification: names of Dependent (Y) and Independent (X) variables

```
=== Model  Y: pltill     X: punemp
```

Regression coefficients: constant in first column; coefficients for each independent variable in subsequent columns, presented in the order in which their names are listed; each row (model) per randomly selected sub-set of user-supplied input areas

r^2: the r^2 associated with the reported regression model (found in final column of row)

```
Expected Regression coeffs; R-square
    7.9616    0.9400   80.7274
   10.6446    0.4752   36.7732
   11.4141    0.3894   10.9613
    8.1415    0.8779   69.7667
    8.6139    0.7912   35.5434
    5.9973    1.2986   69.2406
   11.0467    0.5168   12.4061
    5.0786    1.2359   65.7808
   11.1482    0.5518   39.5141
    9.1323    0.7036   30.4335

 Observed Regression coeffs; R-Square
    8.2322    0.8994   79.1293
   10.6751    0.4648   32.4387
   11.6560    0.3545    9.4335
    8.3563    0.8523   69.1723
    8.6795    0.7745   31.9071
    6.2208    1.2636   65.7610
   10.9679    0.5294   11.4569
    5.4547    1.1838   60.5001
   11.3264    0.5261   36.2878
    9.4537    0.6482   26.9229
```

The third block reports the standard errors associated with the coefficients for regression models fitted to the expected data (to help provide an indication of whether or not the coefficients of models fitted to observed data more than might be expected by chance alone). The first column of

this block represents the standard error of the model constant. Subsequent columns represent the standard errors of the independent variable(s) in the model, listed in the same order as given in the model specification in block 1.

```
SEs of Exp. Regression coeffs
    1.0145    0.1452
    1.3375    0.1970
    2.4361    0.3509
    1.2784    0.1828
    2.1169    0.3369
    1.7102    0.2737
    3.1367    0.4343
    2.1045    0.2819
    1.5237    0.2159
    2.3078    0.3364
```

The fourth block reports the difference (observed – expected) between the regression coefficients and r2 arising due to disclosure control (remembering that the models have been fitted to the same sub-set of user-supplied input areas). The column layout follows that of the first two blocks.

```
 Error [Obs-Exp] of Regression coeffs [& R-square]
    0.2706   -0.0405   -1.5981
    0.0304   -0.0104   -4.3345
    0.2419   -0.0349   -1.5278
    0.2148   -0.0256   -0.5944
    0.0656   -0.0167   -3.6364
    0.2235   -0.0350   -3.4796
   -0.0788    0.0126   -0.9492
    0.3761   -0.0521   -5.2807
    0.1781   -0.0257   -3.2262
    0.3213   -0.0554   -3.5106
```

(vi) *report_full_MR_signs_and_SEs*

Having fitted a series of regression models, SDC_Indirect_Impacts identifies all regression coefficients from models fitted to post-disclosure control data that differ from their pre-disclosure control equivalents by:

(a) sign
(b) more than the standard error associated with the model fitted to pre-disclosure control data
(c) more than the 95% confidence interval associated with the model fitted to the pre-disclosure control data

If this report option is requested, the results of these comparisons are reported in three separate blocks.

The first block reports, for each selected sub-set of user-supplied input areas (rows), and for each regression coefficient (columns), whether or not the pre- and post-disclosure control signs differ (1='yes'; 2 ='no'). The final column in this block reports the total number of regression coefficient signs that differ per fitted regression model:

```
~~ sample Constant  punemp     Total_wrong_signs

Sign_errors
  Count:
          1         0         0         0
          2         0         0         0
          3         0         0         0
          4         0         0         0
          5         0         0         0
          6         0         0         0
          7         0         0         0
          8         0         0         0
          9         0         0         0
         10         0         0         0
```

The second block follows the same format as the first, but identifies and counts those post-disclosure control coefficients that differ by more than one standard error from their pre-disclosure control equivalents.

```
SE_errors
          1         0         0         0
          2         0         0         0
          3         0         0         0
          4         0         0         0
          5         0         0         0
          6         0         0         0
          7         0         0         0
          8         0         0         0
          9         0         0         0
         10         0         0         0
```

The third block also follows the same format as the first, but identifies and counts those post-disclosure control coefficients that fall outside the 95% confidence interval of their pre-disclosure control equivalents.

```
95CI_errors
          1         0         0         0
          2         0         0         0
          3         0         0         0
          4         0         0         0
          5         0         0         0
          6         0         0         0
          7         0         0         0
          8         0         0         0
          9         0         0         0
         10         0         0         0
```

(vii) *report_summary_MR_signs_and_SEs*

This report options provides a summary of the results from *report_full_MR_signs_and_SEs* (above) by reporting the % of coefficients, across all sample sub-sets, that differ in sign (row 1), exceed the expected coefficent value by ±SE (row 2), or fall outside the 95% confidence interval (row 3). The first results column relates to the model constant. Subsequent columns relate to coefficients associated with the model independent variables (in input order). The final column gives the overall % across all coefficients in the model.

```
MODEL:    Constant  punemp     Total_wrong_signs

Sign_errors   % :           0.0000    0.0000    0.0000
SE_errors     % :           0.0000    0.0000    0.0000
95CI_errors   % :           0.0000    0.0000    0.0000
```

(viii) *report_MR_coeff_CIs*

This option produces a report on the distributional range of regression model coefficients and r^2 values, identifying the maximum, minimum, mean, 5th and 95th percentile values for each model when fitted to a series of randomly selected samples (sub-sets) of user-supplied input areas. The report falls into three parts dealing with, respectively, models fitted to pre- and post-disclosure control data, and to the difference in coefficients between the two models. The first results column represents the model constant. Second and subsequent columns represent the independent variable(s) in the model, listed in the order given in the model specification. The final column relates to the model r^2.

```
=== Model  Y: pltill     X: punemp

 Expected Regression coeff (& R-square) CIs
Max.       11.4141    1.2986   80.7274
95%_CI     11.2945    1.2704   75.7951
Mean        8.9179    0.7780   45.1147
```

```
5%_CI       5.4920    0.4280   11.6114
Min.        5.0786    0.3894   10.9613

 Observed Regression coeff (& R-square) CIs
Max.       11.6560    1.2636   79.1293
95%_CI     11.5077    1.2277   74.6487
Mean        9.1023    0.7497   42.3010
5%_CI       5.7995    0.4041   10.3440
Min.        5.4547    0.3545    9.4335

 Regression coeffs (& R-square) Error CIs
Max.        0.3761    0.0126   -0.5944
95%_CI      0.3515    0.0022   -0.7541
Mean        0.1844   -0.0284   -2.8137
5%_CI      -0.0296   -0.0539   -4.8549
Min.       -0.0788   -0.0554   -5.2807
```

(ix) *report_summary_R_outputs*

As well as fitting a series of bivariate and multiple regression models, *SDC_Indirect_Impacts* correlates the user-supplied dependent variable (percentage in first column of user-supplied input percentage data) with all other user-supplied percentages. The results of these correlations, undertaken once for each randomly selected sub-set of user-supplied input areas, is reported if this output option is selected. (Each results row = one sub-set of user-supplied areas).

First the expected (pre-disclosure control) and observed (post-disclosure control) correlation coefficients are reported, then difference between them:

```
=== percentages_v3_P20[popdens]_n20[C]_s100.fmt ===

 Expected correlation coefficients

punemp    pfultim   pownocc   plarent   pnocent   pusdst    pshared   pnocarhh
    0.8985   -0.4308   -0.9381    0.9666    0.6844   -0.6470    0.2647    0.9286
    0.6064   -0.7555   -0.7181    0.6940    0.6092   -0.3209    0.4165    0.6803
    0.3311   -0.7402   -0.7377    0.7505    0.1689   -0.6079    0.0778    0.8654
    0.8353   -0.6236   -0.8923    0.9027    0.8222   -0.4266    0.2016    0.9403
    0.5962   -0.0658   -0.5931    0.6427    0.6252   -0.5476    0.1554    0.7383
    0.8321   -0.6794   -0.8437    0.8575    0.7746   -0.5668    0.2059    0.8657
    0.3522   -0.6479   -0.5953    0.6488    0.5365   -0.4140    0.1167    0.5800
    0.8111   -0.5218   -0.7246    0.7966    0.1525   -0.3928   -0.3387    0.8127
    0.6286   -0.7644   -0.7823    0.7906    0.1525   -0.5240    0.3286    0.7238
    0.5517   -0.6352   -0.8312    0.8179    0.6839   -0.6378    0.1211    0.8491

 Observed correlation coefficients

punemp    pfultim   pownocc   plarent   pnocent   pusdst    pshared   pnocarhh
    0.8895   -0.4348   -0.9435    0.9695    0.6847   -0.6437    0.2057    0.9254
    0.5695   -0.7553   -0.6845    0.6543    0.5791   -0.2575    0.3163    0.6303
    0.3071   -0.7589   -0.7228    0.7515    0.1263   -0.6214    0.1023    0.8512
    0.8317   -0.6568   -0.8984    0.9087    0.8132   -0.4601    0.0980    0.9446
    0.5649   -0.1025   -0.5699    0.6189    0.5820   -0.4886    0.1412    0.6982
    0.8109   -0.6899   -0.8469    0.8536    0.7549   -0.5593    0.0909    0.8580
    0.3385   -0.6571   -0.5665    0.6373    0.4785   -0.3879    0.0734    0.5516
    0.7778   -0.5245   -0.7415    0.8088    0.1816   -0.4348   -0.2833    0.8218
    0.6024   -0.7643   -0.7483    0.7598    0.1077   -0.4859    0.2929    0.6818
    0.5189   -0.6592   -0.8024    0.8067    0.6314   -0.6174    0.0604    0.8245

 Correlation coefficients errors (observed-expected)

punemp    pfultim   pownocc   plarent   pnocent   pusdst    pshared   pnocarhh
   -0.0089   -0.0040   -0.0054    0.0030    0.0003    0.0032   -0.0591   -0.0032
   -0.0369    0.0002    0.0336   -0.0396   -0.0301    0.0634   -0.1002   -0.0500
   -0.0239   -0.0187    0.0148    0.0010   -0.0427   -0.0135    0.0245   -0.0142
   -0.0036   -0.0332   -0.0060    0.0060   -0.0090   -0.0334   -0.1036    0.0043
   -0.0313   -0.0367    0.0232   -0.0238   -0.0432    0.0590   -0.0142   -0.0402
   -0.0212   -0.0106   -0.0032   -0.0039   -0.0197    0.0075   -0.1150   -0.0078
   -0.0137   -0.0092    0.0289   -0.0116   -0.0580    0.0261   -0.0433   -0.0283
   -0.0332   -0.0027   -0.0170    0.0122    0.0291   -0.0420    0.0554    0.0090
   -0.0262    0.0001    0.0341   -0.0308   -0.0447    0.0381   -0.0357   -0.0420
   -0.0328   -0.0240    0.0288   -0.0112   -0.0525    0.0204   -0.0607   -0.0246
```

(x) *report_full_R_signs_and_SEs*

If this option is selected, those correlations reported in the output from *report_summary_R_output* (above) that differ in sign pre- and post- disclosure control are explicitly identified. The layout, although similar to that used in report_summary_R_output, is augmented by table marginals counting the total number of signs that differ with a given sub-set of areas (column marginal 'total') and the total number of signs that differ across all sub-sets of areas for a given percentage (row marginal 'total').

E.g.

Sample	punemp	pfultim	pownocc	plarent	pnocent	pusdst	pshared	Total
	1	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0
	4	0	0	0	1	0	0	1
	5	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0
Total		0	0	0	0	0	0	1

[N.B. figures in this example made up for illustrative purposes]

(xi) *report_summary_R_signs_and_SEs*

The output from *report_full_R_signs_and_SEs* is summarised in this optional output, which records the % of sub-set areas in which the pre- and post-disclosure control correlation signs differ. The final column records the % of all correlation coefficients that differ in sign, across all input percentages and all selected sub-sets of user-supplied areas.

Correlation signs & SE errors

punemp	pfultim	pownocc	plarent	pnocent	pusdst	pshared	overall
0.0000	0.0000	0.0000	0.0000	9.0000	0.0000	0.0000	3.7500

[Figures in this example made up for illustrative purposes]

(xii) *report_R_coeff_CIs*

Reports on the distribution of the coefficients produced by correlating all independent variables in turn with the user-specified dependent variable across all randomly selected sub-sets of user-supplied areas. Distributional measures given are the maximum, minimum, mean, 5th and 95th percentiles. Distributional measures are given separately for correlations observed in pre- and post-disclosure input datasets, and for the difference between pre- and post-disclosure correlation values. The results columns are presented in the same order as independent variables are set-out user-supplied percentage input data.

Expected Correlation coeff. CIs

Max.	0.8985	-0.0658	-0.5931	0.9666	0.8222	-0.3209	0.4165
95%_CI	0.8700	-0.2301	-0.5941	0.9378	0.8008	-0.3533	0.3769
Mean	0.6443	-0.5865	-0.7656	0.7868	0.5210	-0.5085	0.1550
5%_CI	0.3406	-0.7604	-0.9175	0.6454	0.1525	-0.6428	-0.1513
Min.	0.3311	-0.7644	-0.9381	0.6427	0.1525	-0.6470	-0.3387

Observed Regression coeff (& R-square) CIs

Max.	0.8895	-0.1025	-0.5665	0.9695	0.8132	-0.2575	0.3163
95%_CI	0.8635	-0.2521	-0.5680	0.9422	0.7869	-0.3162	0.3058
Mean	0.6211	-0.6003	-0.7525	0.7769	0.4939	-0.4957	0.1098

```
5%_CI       0.3212   -0.7619   -0.9232    0.6271    0.1161   -0.6337   -0.1286
Min.        0.3071   -0.7643   -0.9435    0.6189    0.1077   -0.6437   -0.2833

 Correlation coeff. Error CIs
Max.       -0.0036    0.0002    0.0341    0.0122    0.0291    0.0634    0.0554
95%_CI     -0.0060    0.0001    0.0338    0.0094    0.0161    0.0614    0.0415
Mean       -0.0232   -0.0139    0.0132   -0.0099   -0.0271    0.0129   -0.0452
5%_CI      -0.0352   -0.0351   -0.0121   -0.0357   -0.0555   -0.0381   -0.1098
Min.       -0.0369   -0.0367   -0.0170   -0.0396   -0.0580   -0.0420   -0.1150
```

Program outputs

1) *Screen output*

When run, *SDC_Indirect_Impacts* produces a limited amount of screen output to indicate program progress. These essentially indicate, first, the current input file being processed and, second, the current regression model being processed. In the case of mis-supplied user input, error messages may also appear in the screen output.

2) *SDC_Indirect_Impacts_results.txt*

The main output from *SDC_Indirect_Impacts* is written to this file. The contents of the file will depend on the report options selected (see *SDC_Indirect_Impacts_run_parameters.txt* under 'Program Inputs' above*)*

Appendix: Convert_SAS

CASWEB supplies census data in vector rather than tabular form.

E.g.

```
"ZoneID","s710001","s710002","s710003","s710004","s710005","s710006","s710007","s710008","s710009","
s710011"
"04BXFA01 ",458,0,453,5,134,2,50,0,50,13
"04BXFA02 ",280,0,272,8,81,2,24,0,24,8
"04BXFA03 ",706,0,697,9,200,2,69,0,69,21
"04BXFA04 ",446,0,439,7,200,6,50,0,50,20
```

Note that:

- these vectors are supplied in comma-separated format, one row per area
- each vector starts with an area name
- the first row of the file comprises a set of 'headers' describing the vector contents
- all string variables (area names and headers) are placed between quotation marks (e.g. "04BYFA32").
- Only table cells that can have a valid count are included in CASWEB vectors (hence no header/count for SAS Table 71, cell 12)
- The maximum vector size supported by CASWEB is 240 cells; for tables larger than this, data have to be extracted as a pair of complementary vectors.

The program *ConvertSAS* takes CASWEB supplied vectors and for each table of interest:

(i) Splices together multiple vectors (if necessary)
(ii) Adds back missing vector cells (setting their value to zero)
(iii) Converts vector to tabular format
(iv) Writes out table, omitting any rows/columns as required
(v) Creates an additional file, *threshold.fmt*, based on SAS Table 71, that reports the number of residents and households present in each area processed.

The results from this conversion procedure are suitable for use as input to *Perturb_v2*, and are written to the folder *Convert_SAS_outputs*.

The folder *Convert_SAS_inputs* should be used to store the inputs to *Convert_SAS*. These inputs comprise the *csv* format CASWEB-supplied table-specific vectors, plus two control files that identify:

- any vectors that require splicing;
- no. of rows and columns in each table
- rows and/or columns to be omitted from program output
- table cells omitted from CASWEB-supplied vectors

Control file 1: *List_of_table_parts_to_splice.txt*

For each table supplied as two vectors, the following information is required:
Name of input vector 1; Name of input vector 2; Name of file for output of combined vector; no. of cells in first vector (excluding area name); no. of cells in second vector (excluding area name)

e.g. `"Liv_SAS08a.csv","Liv_SAS08b.csv","Liv_SAS08.csv",240,58`

Control file 2: *List_of_inputs_for_ConvertSAS.txt*

For each table supplied as a vector, the following information is required:

- Input vector filename (post vector splicing, if any)
- Output table filename
- Indicator of whether or not table is based on splicing two vectors (1=yes; 0=no)
- No. of cells in SAS table
- No. of cells missing from CASWEB vector
- The Cell numbers of any missing cells
- No. of columns in SAS table
- No. of rows in SAS table
- No. of columns to be omitted from table during output
- List of columns to be omitted from table output (if any) (columns numbered from left to right; left-most = 1)
- No. of rows to be omitted from table during output
- List of rows to be omitted from table output (if any) (rows numbered from top to bottom; top row = 1)

e.g. `"Liv_SAS71.csv","table71.fmt",0,12,2,10,12,6,2,0,0`