CASC PROJECT

Computational Aspects of Statistical Confidentiality

September 2003

TM1 - The Evaluation of Risk from Identification Attempts

Dr. M. J. Elliot

Dr. K. Purdam

CCSR
University of Manchester
Oxford Road
Manchester M13 9PL

Deliverable No: 5-D3

CASC Project Deliverable 5D3:

TM1 - The Evaluation of Risk from Identification Attempts

Dr. M. J. Elliot
Dr. K. Purdam

SEPTEMBER 2003

CCSR
University of Manchester
Oxford Road
Manchester M13 9PL
Tel: 0161-275 4721
Fax: 0161-275 4722

CONTENTS

1.Introduction

This paper summarises findings from research into the risk from identification attempts from matching information available in the public domain with data released by national statistic institutes.

The research builds on the evaluation of the availability of data sources that could be used for identification purposes in the EU, the findings of which are available in Elliot and Purdam (2002). This research also develops earlier work, which attempted to match household and census records and which yielded very valuable results, in particular highlighting the protection that arises from highly correlated categorical variables.

Using the UK General Household Survey and a sample survey data from a local authority we have attempted to mimic an intruder seeking to establish identification. Starting from two or three defined scenarios to define key variables we then assessed the availability of matching data We compiled an identification database from a range of publicly available and restricted access databases and conducted a series of matching analyses. Data sources for matching attempts included occupational registers, electoral registers, GP lists, housing information.

The Office for National Statistics in the UK assisted in validating the matches/identification on the anonymised GHS data. The results highlight the weakest points in protection and thus indicate the particularly risky variables or combinations of variables. The research also assesses the degree to which identification is impeded by the application of the disclosure control methods implemented in ARGUS. The results allow estimates of identification given a high level of resource input.

The work employed the new "Data Intrusion Simulation" method recently developed under funding from the UK Economic and Social Research Council and the US Bureau of the Census, which provides estimates of the probabilities of correct matching against a give target file.

2. Data Collection

1. PUBLICLY AVAILABLE DATA

The examination of publicly accessible data sources revealed a wide range of information available on individuals. The identification variables were initially specified in an earlier stage of the research, the findings of which are available in Elliot and Purdam (2002).

It is clear that there has been a substantial growth in the collection, storage and release of personal data across the public and private sector in Europe. In addition, far more personal information is collected and kept on restricted access databases across the private public and voluntary sector. Evidence suggests that the confidentiality practices in place around such data sets vary considerably.

Across Europe, public records and information available from the Internet contain a wide range of personal information. New databases and types of data are being constructed and made available each day, such as, for example, databases of personal communication and movement. The increased number of sources of personal data results in an increasing number of details on each individual being stored. Linking data within and across organisations is a major issue and data is increasingly transmitted and shared across national borders.

2. TARGET DATA

General Household Survey Data

The General Household Survey (GHS) in the UK is a multipurpose household survey conducted by the ONS. It was begun in 1971 and is carried out annually on a sample of 13,000 addresses taken from the Postcode Address File. All adults aged 16 and over are interviewed in each responding household. The GHS collects information on over 600 variables including date of birth, gender, marital status, housing type and tenure, consumer durables, employment, education, income, migration, health (incl. smoking and drinking), care and family information. It has specific sections for particular types of individuals such as the elderly.

Samples of Anonymised Records

The Individual level Samples of Anonymised Records were a 2% sample of 1991 British census data. The data contains 45 demographic, economic, health and housing variables and is released under license to UK academics, local authorities and other institutional users.

Local Authority Sample Survey Data

It was also considered important to examine the identification risks in relation to restricted access databases. Such data sets are widely produced in the UK often down to very local levels. Though access to such databases is not public evidence suggests such databases are not always held under confidential conditions. In addition, the data has

often been collected via a third party sub-contractor that can also pose identification opportunities to an intruder.

With the agreement of a local authority a sample household survey data set was obtained for the purposes of the matching experiments. The data set was provided under strict confidential terms of use. The data set contained detailed demographic information of over 800 individuals within a local area. Key variables included: name, address, age, gender, household type, tenure, no of rooms, no of children, no of cars, employment, garage, health, smoking etc **KP to check and add.** A number of questions were designed specifically for the survey while others were taken from existing national surveys on the basis of developing comparable baseline information.

3. IDENTIFICATION DATA

Neighbourhood observation data

Using a combination of publicly available data from the electoral register and sources such as local newspapers, estate agents, the Internet and a range of observation techniques an identification data set was compiled. Key variables included: name, address, house type, no of rooms, no of cars, vans, motorbike, caravans, garage, double glazing, satellite TV, burglar alarm. In addition observable data in terms of children in the household or elderly people was also connected when available.

If the house of a similar house was for sale observation details were confirmed with details from estate agents where available.

Commercial lifestyle data

A comprehensive lifestyle database of the area was also purchased. Such data sets are widely available in the public sphere at low costs. Key variables purchased include: name, address, post code, gender, age, income, occupation, no of children, house type, tenure and length of residence.

Other variables were available. See Appendix 1 for an example list.

Neighbourhood Statistics www.neighbourhood.statistics.gov.uk/

Data from the Office for National Statistic neighbourhood statistics service were also used. The Neighbourhood Statistics Service offers ready access to a vast range of social and economic aggregate data relating to a consistent small-area geography. Data includes: population demographics, education and training, employment, housing, health care and access to services.

3. The Experiments

3.1 General Method

Previous work has shown us (Elliot and Dale 1998) that the initial problem was to find the best method for distinguishing false from true matches. Fortunately, in parallel with this project Elliot et al (2002) have developed the special uniques identification algorithm, which is a method that uses information available in a microdata file to identify risky records. Recant advances by Elliot and Manning (2003) have shown how the method can be used to effectively grade the matches in terms of riskiness. By definition one can tie this method into the matching algorithm it is possible to distinguish between low and high probability matches.

The method then was multi layered.

1) **Data Preparation:** Extensive data preparation was carried out in order to optimally align the files for matching.
2) **Simple Key Variable Matching:** Key variable matching using a simple algorithm was carried out. The number of correct matches was recorded.
3) **Refined Matching:** The matches were graded using the special uniques algorithm. The ten/fifty most probable matches were chosen. The proportions of correct matches were recorded.
4) **Application of Argus:** Argus was applied to the Target set.
5) **Refined Fuzzy matching** in combination with the special uniques algorithm is used to attempt to recover the matches. The ten/fifty most probable matches were chosen. The proportions of correct matches were recorded.

3.2 An Attack on a Government Microdata Set.

To simulate is an attempt to link records from an outside database with records in a government microdata file, we utilised earlier work simulating an attempt to Link two British datasets. The "target" microdata set was the 2% individual Sample of Anonymised Records from the 1991 Census and the 'identification file' was a subset of variables related to health, drawn from the General Household Survey (GHS) for 1991. Although these datasets are specific to Britain, the results obtained from this experiment can be generalised to any surveys that collect comparable data.

Previous work that has attempted to link records between databases has mainly been conducted in the field of Record Linkage and has been concerned with issues such as linking similar strings (for example, names, some of which may be mispelt), string comparators (see for example Winkler 1994) and the optimisation of expectation maximisation (EM) algorithms. However, much concern in disclosure control stems from concerns about matching released microdata file with another dataset held by an intruder. The level of success that can be achieved with this type of record linkage has been much less extensively studied, with the notable exception of the experiment by Muller et al (1992). Therefore in this experiment we set out to assess how accurately a match can be achieved in circumstances where we know the expected overlap between two databases.

3.2.1 Preparing the data sets

In order to conduct the experiment, it was first necessary to translate the data on both datasets into a mutually coherent form of standard key variables. This was by no means a trivial task. We had, however, assumed that because the datasets were derived from the same source they would relatively easy to harmonise. In fact, even for variables relating to the same information and apparently categorised in the same way, the observed frequencies were often quite different.

Data Ageing.

The GHS dataset covered 12 months from April 1991 to March 1992. Therefore the data may have been collected from two weeks before the 1991 Census date to eleven and a half months after census date. Initially, this was seen as a problem to be overcome in the matching algorithm. However, as we shall see later, the number of matches was so large that it was decided to focus solely on the data from April 1991 (census month). This minimised the data ageing effect and increased the probability of successful unique matches.

In all 17 standard key variables were used for matching between the two sets these were:

- age – 94 categories
- sex
- marital status – 5 categories
- country of birth – 22 categories
- economic activity – 11 categories
- level of educational qualification – 4 categories
- family type – 4 categories
- presence of a long term limiting illness
- ethnic group – 10 categories
- migration in last year – 3 categories
- socio-economic group of those in employment – 17 categories
- housing tenure – 6 categories
- whether head of household has a long term limiting illness
- presence of a dependent child in the household
- presence of pensioners in the household
- number of residents in the household – 4 categories
- number of cars – 4 categories

The full translation of these variables from the GHS and the SARs to standardised key variables (SKV) is shown in appendix A to this report. However, it worth noting here two particular features of the GHS dataset that differ from a normal identification file.

Age: The lack of a date of birth variable on the GHS file means that exact age is not known. Thus age recorded in the SAR file may be either 1 year less than or the same as the age given in the GHS file.

Geography: The lowest geography on the GHS dataset is a twenty-two point sub-regional geography. This is a much cruder geography than the areas on the 2% SARs. Although these two files do not provide the exact matching keys – such as date of birth or names – that may be available in many record linkage experiments, they maximise the chances of matching in other respects. For example, both datasets are collected by the same organisation; there is a great deal of similarity in the questions asked for each variable; the timing of data collection can be matched very closely.

3.2.2) Results of key variable matching

When records from the GHS were matched against the SARs using the seventeen variable key about 50% of records in the GHS matched against one or more individual in the SARs. In many cases there were very large numbers of 'statistical twins' in the SARs for one GHS record. Clearly this made the matching algorithm developed in the last section redundant, since it only has value in the context of prioritising partial matches. With the large number of exact matches, the algorithm has no value.

To reduce the task to manageable proportions we focused on one GHS collection month - April 1991 - which was the month of census enumeration and would therefore ameliorate the effect of data ageing. On the basis of the sampling fraction, about 40 records were expected to occur in both samples for April.

Our matching system identified 219 records in the April 1991 GHS file which matched one, and only one, individual in the SARs and a further 112 records which matched 2 individuals in the SARs. This is the point to which any intruder could proceed and it was impossible to narrow down these matches any further.

In order to establish how many of these apparent matches were correct the two files were sent to ONS for verification.

Verification procedure

The output from the matching experiment was in the form of a Microsoft Excel file containing those cases that corresponded to matched pairs. We entered the first six fields in the output file, shown below, were entered using the released data and the file was then sent to ONS Social Survey Division who added date of birth and the address fields. The file was then passed to ONS Census Division who completed the match check field indicating a true/false match. They then deleted the data in the date of birth and address fields before returning the file to the experimenter [1]

GHS house number
GHS person number
SAR ID number
GHS HMONTH
GHS HYEAR
Sex
Date of birth
ADDRESS
MATCHCHECK

As a result of this checking procedure ONS were able to confirm that six of our 219 unique matches were correct and that there were a further two matches in the 112 records with two possible matches.

This is largely explained by the fact that most of the variables used for matching are categorical and there is a high degree of inter-dependency between them. For example, children under 16 will all take the same codes on marital status, economic activity,

[1] Note: because there are no real world identifiers on either of the original files there is no possibility of unwanted actual identification in this procedure.

qualification level, presence of a dependent child in the household. Most will also be white, living with two parents and with no long term limiting illness. Similar dependencies will be present for pensioners and for the unemployed. Even amongst the employed people in the sample there were high levels of correlation.

The DIS method (see Skinner and Elliot 2003) was run on the SAR with same key variables and level of geographical detail as was used in the matching experiment. This established the theoretical probability that a unique match was a correct match, asuming no data divergence. The estimated probability using this method was 0.088. This is equivalent to 19.26 correct unique matches if replicated in the GHS file. This indicates a *data divergence rate, for one or more variables* between the two files of:

$$1- (6/19.26) \approx 0.68$$

So the DIS method implies that even before ARGUS has been applied to the SARs

3.2.3 Application of SUDA

The latest version of SUDA was employed (see Manning and Elliot 2003), only the matches against uniques were considered. This enabled the matches to be coded according to how probable they were to be correct (if there were no data divergence). The top ten/fifty matched records were selected. These had a mean matching probability of 0.66/0.40 respectively. Given that the DIS result suggested a divergence rate of about 0.68, we would expect a real match rate of 0.22/0.13.[2] In fact two of the top ten (20%) matches were correct and four of the top fifty (8%), indicating the SUDA algorithm was working basically as expected and has honed the effective match rate (although the errors on these figures are potentially huge).

3.2.4 Application of Argus

The decision about how to conduct the perturbation study was one of the more problematic aspects of this work. ARGUS is a disclosure control tool rather an automated disclosure control system. As such it leaves decisions about key variable combinations and parameter selection to the user. As the data we were using was unweighted, the risk model did not apply and therefore we had no internal to Argus means of making decisions on the basis of levels of risk. A further problem was that it was not possible to use full scenario based keys (as developed by Elliot and Dale 1999).[3]

For consistency with other work on this data set (Elliot and Manning 2003), various combinations were produced experimentally and programs extracting the records and variables thus identified as risky were compared with the outputs of the special uniques program. the following variable combination frames appeared to identify risk in a similar way to that program:

A. All individual variables (threshold=4)
B. All pairs of variables (threshold=2)

[2] Assuming that the data divergence is random in relation to special uniqueness.
[3] A further more general problem with ARGUS is that it is not possible to block missing values for use in perturbations. This means that inconsistencies are produced where not applicable categories are used to record suppressions or as Post randomisation categories.

C. All 3-way combinations under scenarios (threshold=1).

Three perturbed SAR datasets were then produced.

File A. Suppression based file.

On this file the disclosure control was entirely based around suppressions. All three combination levels were used, to determine the suppressions. The default suppression weights were employed.

File B. *PRAM FILE*

All variables on file PRAMed. The per value change probabilities of PRAM were set to maintain the univariate distributions.[4] For some variables such as age bandwidths were used partly to control the number of inconsistencies.

File C. *Combined Pram and suppressions.*

Suppressions were applied to the PRAM FILE C, with only level A and B combinations being used.

3.2.5 Results of Matching after Disclosure Control

The matching algorithm was run again on each of the three perturbed files. The results of the key variable match are shown in table 1. Using the SUDA algorithm to refine the key variables gives the results in Tables 2 and 3.

Table 1: The impact of ARGUS on the level of key variable matching.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A	207	3	0.014	-47%	-50%
B	223	2	0.009	-67%	-67%
C	179	1	0.006	-80%	-83%

Table 2: The impact of ARGUS on the level of SUDA refined key variable matching top 10.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A	9	1	0.1	-50%	-50%
B	9	1	0.1	-50%	-50%
C	9	1	0.1	-50%	-50%

Table 3: The impact of ARGUS on the level of SUDA refined key variable matching top 50.

[4] When using PRAM, it would be useful if it were to control more exactly the change matrix probabilities than simply through bandwidths.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A	47	3	0.06	-25%	-25%
B	48	2	0.04	-50%	-50%
C	48	2	0.04	-25%	-25%

The method of fuzzy matching we used is similar to that used in record linkage work (see for example Winkler 1994, Felligi and Sunter 1969). The method is simpler than that used in record linkage work for three reasons:

1. except for age all matching variables are either categorical or crudely ordinal. There is little or no possibility of *degrees* of data divergence for a particular value.
2. The overlap between the data sets is small (at for example ~0.04% of the INDSAR), so an algorithm to resolve competition between potential matches is superfluous.
3. As we did not link identification files we are not concerned with typographical/spelling errors in text (such as name/address).

(i) In line with Winkler's (Newcombe's 1988) method a four variable blocking key is used, these areage, sex, marital status and region. For each combination of these variables a file is produced for the INDSAR. This produces 18,800 files with and average ~650 persons in each.

(iii) For each of the monthly GHS datasets, the following procedure is conducted:

(a) A set of active objects is constructed. Each object corresponds to one GHS record. In each run (corresponding to a single months data from GHS) ~2000 such objects will be created.

(b) EACH object attempts to match itself against each record in the two appropriate blocking key files (one file for each of two possible ages). It produces a value for itself referred to as it's activation energy, which represents the value of the best possible match from the blocking key files. These values are determined by the following:

$$\frac{w_1v_1 + w_2v_2 + \ldots + w_jv_j}{\Sigma w}$$

v is a real number between 0 and 1 which in the current experiment has one of three values

1 = direct match
0.5 = match from/to a wild-card (or match within 1 year for age)
0 = no match

for each give key variable from 1 to j.

w (1-j) are real numbers between 0 and 1 being weights associated with each variable which reflect the reliability of the variable, in terms of likely data divergence.

If having evaluated itself an object has an energy value of less than a threshold value then this object drops out of the set of possible matches. If two or more objects have best possible matches with the same blocking file record then the one with the highest energy is presumed and the one with the lower energy re-evaluates itself to it's next best match.[5] This procedure continues until objects have a single match (or a set of equal matches).

Table 2: The impact of ARGUS on the level of SUDA fuzzy matching top 10.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A					
B					
C					

Table 3: The impact of ARGUS on the level of SUDA refined fuzzy matching top 50.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A					
B					
C					

3.3 An Attack on a Local Authority Database

We were fortunate to obtain access to a local authority survey dataset for one ward with a UK local authority to use as a target set.[6] The data was a survey of households collected to test the impact of a UK government initiative but resulted a fairly rich dataset containing demographic, economic and health indicators. The identification dataset was a lifestyle dataset bought from CACI UK and was database of 33% of adults within a larger part of the same local authority area including that contained within the target set.

3.3.1 Preparing the data sets

As with the first experiment, it was first necessary to translate the data on both datasets into a mutually coherent form of standard key variables.

Data Ageing

The two datasets were collected at different times some 6 months apart, however as both datasets included a length of residence variable we were able to check for obvious matches as a precursor to the simulation.

[5] This aspect of the procedure is more useful when one of the two datasets is a population dataset or the overlap is much larger than in the current experiment.

[6] The data was held under special license and under CAPRI standard high security conditions. The data was held for the duration of the study and then destroyed.

In all nine standard key variables were used for matching between the two sets these were:

- age
- sex
- marital status
- occupation
- housing type
- number of children
- length of residence
- housing tenure
- number of cars

3.3.2) Results of key variable matching

Verification for this experiment was considerably simpler than experiment 1, as both datasets had identification information present. The data sets were sufficiently small to make its possible to completely clean name and address information before the matching study was conducted. Checking of matches was carried out manually.

<INSERT RESULTS>

The DIS method (see Skinner and Elliot 2003) was run on the LAD with same key variables and level of geographical detail as was used in the matching experiment. This established the theoretical probability that a unique match was a correct match, assuming no data divergence. The estimated probability using this method was ????. This is equivalent to ??? correct unique matches if replicated in the LIF file. This indicates a *data divergence rate, for one or more variables* between the two files of:

3.3.3 Application of SUDA

The latest version of SUDA was employed (see Manning and Elliot 2003), only the matches against uniques were considered. This enabled the matches to be coded according to how probable they were to be correct (if there were no data divergence). The top ten/fifty matched records were selected. These had a mean matching probability of ???/??? respectively. Given that the DIS result suggested a divergence rate of about??? we would expect a real match rate of ???/???.[7] In fact ??? of the top ten (?%) matches were correct and four of the top fifty (?%).....

3.2.4 Application of Argus

For consistency with experiment 1 the following variable combination frames appeared to identify risk in a similar way to that program:

D. All individual variables (threshold=4)
E. All pairs of variables (threshold=2)
F. All 3-way combinations under scenarios (threshold=1).

[7] Assuming that the data divergence is random in relation to special uniqueness.

So again three perturbed LAD datasets were then produced.

File A. Suppression based file: On this file the disclosure control was entirely based around suppressions. All three combination levels were used, to determine the suppressions. The default suppression weights were employed.

File B. *PRAM FILE:* All variables on file PRAMed. The per value change probabilities of PRAM were set to maintain the univariate distributions. For some variables such as age bandwidths were used partly to control the number of inconsistencies.

File C. *Combined Pram and suppressions:* Suppressions were applied to the PRAM FILE C, with only level A and B combinations being used.

3.3.5 Results of Matching after Disclosure Control

The matching algorithm was run again on each of the three perturbed files. The results of the key variable match are shown in table 5.

Table 6: The impact of ARGUS on the level of key variable matching.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A	207	3	0.014	-47%	-50%
B	223	2	0.009	-67%	-67%
C	179	1	0.006	-80%	-83%

Table 7: The impact of ARGUS on the level of SUDA refined key variable matching top 10.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A	9	1	0.1	-50%	-50%
B	9	1	0.1	-50%	-50%
C	9	1	0.1	-50%	-50%

Table 8: The impact of ARGUS on the level of SUDA refined key variable matching top 50.

Perturbation file	False Unique Matches	True Unique Matches	Proportion	% Change in Proportion	% Change in Matches
A	47	3	0.06	-25%	-25%
B	48	2	0.04	-50%	-50%
C	48	2	0.04	-25%	-25%

References

Briggs, M. (1992): *Estimation of the proportion of Unique PopulationElements Using a Sample*. Statistics Canada Working Paper.

Carter, R., Boudreau, J-R, and Briggs, M. (1991) Analysis of the risk of disclosure for census microdata, mimeo, Statistics Canada

Elliot, M.J. and Dale (1998) A Disclosure Risk Matching Experiment Report to European UnionDM 1.5. ESP/ 204 62/DG III/DM.15

Greenburg, B., and Voshell, L. (1990*) The Geographic Component of Disclosure Risk for Microdata,* American Bureau of the Census SRD Reassert Report Census/SRD/RR-90/13

Blien, U., Wirth H. and Muller M. (1990) Identification Risk for Microdata Stemming from Official Statistics. *discussion paper presented at the International Symposium on Statistical Disclosure Avoidance*, Den Haag, Netherlands.

Muller, W; Blien, U.; Wirth, H.(1992): Disclosure risks of anonymous individual data. Paper presented at the 1st International Seminar for Statistical Disclosure.

Skinner, C. J. and Elliot, M. J. (2002). 'A measure of disclosure risk for microdata', Journal of the Royal Statistical Society Series B, 64(4) pp 855-867.

Appendix A Commercial data variables

Category	Variable
Gender	Female, Male
Age Band	18-24, 25-34, 35-44, 45-54, 55-64, 65+
Household Composition	Single household, family household, multi-occupancy single sex household, multi-occupancy mixed sex household, pseudo family
Length of Residence	0-2 years, 3-5 years, 6-10 years, 11+ years
Children	0 children, 1-2 children, 3+ children, children aged 0-4, children aged 5-10, children aged 11-15, children aged 16+, attainer/young person in household
Lifestyle Triggers	Empty Nester household, got first job in last year, got married/started living together in last year, had a child in last year
Occupation	Craftsman/tradesman, factory/manual, housewife/husband, medical/education, middle management, office/clerical, professional/senior manager, retired, self-employed, shop worker, student, work in public sector
Qualifications	A levels, O levels/GCSE/none, NVQ/OND/HND, first degree, higher degree/professional,.
Income	Family income £0-£9,999, family income £10,000-£19,999, family income £20,000-£29,999, family income £30,000-£39,999, family income £40,000-£49,999, family income £50,000+, pays no income tax, pays basic rate income tax, pays higher rate income tax
Home Ownership	Owned, rented from council, rented from housing association, rented privately
House Type	1 bedroom, 2 bedrooms, 3 bedrooms, 4 bedrooms, 5+ bedrooms, bungalow, detached, flat/maisonette, semi-detached, terraced
House Value	£0-£60,000, £60,000-£120,000, £120,000-£200,000, £200,000-£500,000, £500,000+
In the Home	Have an office at home, have central heating breakdown cover, have dishwasher, have microwave, have tumble dryer or washer/dryer
Household Technology	Have cable TV, Have digital TV, have satellite TV, subscribe to cable TV or cable phone, use digital TV to make purchases / bookings, have DVD player, have games console with access to internet, have digital camera, have internet access via TV, have video
Telecoms	Have mobile phone, have WAP phone, International calls regularly, regular use of landline telephone, telephone quarterly bill £0-75, telephone quarterly bill £76-99, telephone quarterly bill £100-150, telephone quarterly bill £151+
Computers	Have palmtop computer, have laptop computer, have Mac / iMac, have PC, use home PC for careers / job planning, use home PC for education / reference, use home PC for personal finance, use home PC for home shopping, use home PC for playing computer games

Internet	Use Internet for email, use Internet for shopping : books / cds, use Internet for shopping: clothing/fashion, 1-3 Internet purchases in last year, 4+ Internet purchases in last year, use Internet to buy gifts, use Internet to buy/research cars, use Internet to make leisure and holiday bookings, no Internet purchase in last year
Utilities	Have changed electricity supplier, have changed gas supplier, have changed telephone supplier, have mains gas supply
Holidays	Camping/caravanning, Europe/Med, hotel/hotel package, rest of the world, self-catering, UK/Ireland, USA/Canada, weekend break, winter snow, winter sun
Interests	Angling, bingo, birdwatching, charity/voluntary work, cinema, cookery, current affairs, DIY, eating out, environment/wildlife, exercise/sport, fashion/clothes, fine arts/antiques, football, football pools, foreign travel, gardening, golf, gourmet food/Wine, hiking/walking, home computing, horseracing, listening to music, magazine subscriber, music - classical/opera, music - easy listening, Music - eighties, music - light classical, music - rock and roll, National Trust, photography, pub, reading books, reading historical works, religious activities, rugby, self improvement/education, sewing/needlecrafts, snow skiing, theatre/arts, wines by mail order
Lifestyle	Regularly eat evening meal in pub/restaurant, regularly eat lunch in pub/restaurant, visit coffee bar 3+ times per week, family uses herbal medicines/health foods, family uses vitamins/minerals/supplements, eats brown/wholemeal/granary bread, buyer of environmentally friendly/recycled products, wine buyer (6+ bottles per month), wear contact lenses, wear spectacles
Newspapers - Daily	Daily Express, Daily Mail, Daily Mirror, Daily Record, Daily Sport, Daily Telegraph, Financial Times, Guardian, Independent, Star, Sun, Times
Newspapers - Sunday	Independent on Sunday, Mail on Sunday, News of the World, Observer, Sunday Express, Sunday Mirror, Sunday People, Sunday Sport, Sunday Telegraph, Sunday Times
Groceries - Store	Use home delivery service for weekly shopping, shop at Asda, shop at Morrisons, shop at Safeway, shop at Sainsbury, shop at Somerfield, shop at Tesco
Groceries - Supermarket Spend	Spend £0 to £25 per week, spend £25 to £44 per week, spend £45 to £59 per week, spend £60 to £74 per week, spend over £75 per week
Groceries - Travel to Store	Car, public transport, taxi, walk
Groceries - Reasons for Choice of Store	Children's/creche facilities, customer service, distance, food range, parking facilities, prices, quality of products, store loyalty card
Shopping - Mail Order	Mail order 6+ times per year, catalogue spend £500+ in last year, mail order never
Motoring	Number of cars 0, number of cars 1, number of cars 2+, bought main car new, car insurance £0-300, car insurance £300-500, car insurance £500+, company car, company car user/chooser, has car less than 3 years old, has motorbike, has scooter, keeps main car in garage, keeps main car on driveway/road, likely to spend £0-5,000 on main car, likely to spend £5,000-£10,000 on main car, likely to spend £10,000-20,000 on main car, likely to spend £20,000+ on main car
Finance - Banking	Have current account, Internet account with e-bank, Internet account with traditional bank, switched current account in last year, have telephone banking, consider banking by interactive TV, consider banking by mobile phone, consider PC banking, consider telephone banking

Finance - Cards	Have credit card, have Amex/Diners card, believes credit limit is sufficient for needs, credit card limit £0-999, credit card limit £1,000-£2,499, credit card limit £2,500-£4,999, credit card limit £5,000+, dissatisfied with some aspect of credit cards, have credit card with UK bank, have credit card with UK new player, have credit card with US player, interested in gold or platinum card, monthly credit card spend £0-50, monthly credit card spend £51-100, monthly credit card spend £101-250, monthly credit card spend £250+, new credit card in last year, have retail store card
Finance - Savings	Have savings account, have instant access savings account, have Internet savings account, have National Savings account/investments, have restricted access savings account, have savings account with bank, have savings account with converter, have savings account with new player, have savings account with supermarket, have child savings plan, have guaranteed income bonds, have high interest investments, have regular savings plan
Finance - Investments	Have ISA, have Cash ISA, have Maxi ISA, have Mini ISA, have stocks and shares ISA, have lump sum investment, have stocks and shares, have Unit Trust, arranged ISA directly, arranged ISA through IFA, arranged Unit Trust directly, arranged Unit Trust through IFA
Finance - Mortgages	Have a mortgage, 0-10 years left on mortgage, 11+ years left on mortgage, arranged mortgage directly, arranged mortgage through IFA, have re-mortgage, re-mortgaged with different lender, re-mortgaged with same lender
Finance - Loans	Have a loan, foresee need for personal loan, took out loan for consolidation, took out loan for home improvements, took out loan for new car, took out loan for other spending
Finance - General Insurance	Have general insurance, arranged general insurance directly, arranged general insurance through IFA, have home contents insurance, renewed general insurance with same supplier in last year, wwitched supplier of general insurance in last year
Finance - Health Insurance	Company pays for PMI, considering health insurance, health insurance with BUPA, health insurance with Norwich Union healthcare, health insurance with PPP, health insurance with Prime Health, personally pay for PMI
Finance - Insurance	Have accident insurance, have mortgage protection, have motor insurance, have travel insurance, have funeral plan
Finance - Life Assurance	Have life assurance policy, arranged life assurance directly, arranged life assurance through IFA
Finance - Pensions	Have company pension scheme, have private personal pension scheme, arranged private personal pension directly, arranged private personal pension through IFA
Finance - Channel	Use Internet to buy financial services, use Internet to arrange personal loan, use Internet to buy insurance, use Internet to buy investments/ISAs, use Internet to source credit card, prefer adviser as financial channel, prefer branch as financial channel, prefer broker as financial channel, prefer direct mail as financial channel, prefer Internet as financial channel, Prefer phone as financial channel
Finance - Attitudes	Always knows how much is in bank account, better off having what you want now, dislikes borrowing, dislikes going into branch, regularly reads financial pages, saves only for specific purposes, would be happy to use Internet for banking, would be happy to use phone for banking, would only consider 1 or 2 financial institutions, consider professional financial help
Major Purchases	Intend to finance purchase with credit, intend to finance purchase with savings, plan to spend money on a car, plan to spend money on consumer durables, plan to spend money on a holiday, plan to spend money on home improvement, plan to spend money to pay off debt, plan to spend money on wedding

Charity Concerns	Animal welfare, children, disabled, disaster relief, elderly, environment, medical, third world, wildlife, contribute by covenant or direct debit

Appendix B: Specification relation between GHS variables and Standard Key Variables.

SEX2<-SEX

	Standard Key Variable	GHS Variable
Value Label	SEX 2	SEX
Male	1	1
Female	2	2

AGE94<-AGE

	Standard Key Variable	GHS Variable
Value Label	AGE94	AGE
0-90	0-90	0-90
91	91	91
92	91	92
93	92	93
94	92	94
95 or older	93	95-99

COBIRTH22<-COB

	Standard Key Variable	GHS Variable
Value Label	COBIRTH22	COB
England	1	01
Scotland	2	02
Wales	3	03
Northern Ireland	4	04
UK Other	5	05
Irish Republic	6	06
EEC	7	07
Other Europe	8	08
Old Commonwealth	9	09
India	10	10
African Commonwealth	11	11,12
Caribbean Commonwealth	12	13
Mediterranean Commonwealth	13	14
Far East Commonwealth	14	15
Other New Commonwealth	15	16
Pakistan	16	17
Bangladesh	17	18
Rest of Africa	18	19
America	19	20
Middle East	20	21

Rest of Asia and Oceania	21	22
Elsewhere, Vague response	22	23

PRIMECON11<-ECSTILO+WKSTATE+SEGE

		GHS variables			
Value Label	PRIMECON 11	ECSTILLO	SELFEMPE	WKSTATE	SEGE
Child under 16	0	-6	-	-	-
Employee Full-time	1	1	1	1	-
Employee Part-time	2	1	1	2, 3	-
Self-employed with employees	3	1	2	-	1, 7, 16, 18
Self-employed without employees	4	1	2	-	3, 5, 15, 17
Government Scheme	5	2,3	-	-	-
Unemployed	6	4,5	-	-	-
Student	7	9	-	-	-
Sick	8	6	-	-	-
Retired	9	7	-	-	-
Other Inactive	10	8,10	-	-	-

QUALLEVEL4<-EDLEV

	Standard Key Variable	**GHS Variable**
Value Label	QUALEVEL4	EDLEV
No post 18 qualifications	0	6-17,-9,-6
Post 18 sub degree	1	3-5
First Degree or equivalent	2	2
Higher Degree	3	1

FAMTYPE4<-FUT20

	Standard Key Variable	**GHS Variable**
Value Label	FAMTYPE4	FUT20
Not in family	0	1
Couple no children	1	2
Couple with children	2	3
Lone parent with children	3	4, 5

LTILL2<-LIMILL

	Standard Key	**GHS Variable**

	Variable	
Value Label	LTILL2	LIMILL
Has long term limiting illness	1	1
Has no long term limiting illness	2	2, -9

MARCONA5<-MARSTAT

	Standard Key Variable	GHS Variable
Value Label	MARCONA5	MARSTAT
Single	1	3
Married	2	1,6
Cohabiting	3	2
Divorced	4	5
Widowed	5	4

ETHNIC10<-ORIGIN

	Standard Key Variable	GHS Variable
Value Label	ETHNIC10	ORIGIN
White	1	1, 11, 49-52, 56, 69-72
Black Caribbean	2	6, 16, 22, 42, 63
Black African	3	7, 17, 24, 45, 65
Black Other	4	21, 28, 29, 31
Indian	5	2, 12
Pakistani	6	3, 13
Bangladeshi	7	4, 14
Chinese	8	5, 15
Other Asian	9	25-27, 46-48, 66-68
Other - other	10	23, 30, 41, 42, 44, 53-55, 57, 61, 62, 64, 73-77

MIGRANCY3<-RESLEN5+AGE

	Standard Key Variable	GHS Variable	
Value Label	MIGRANCY3	RESLEN5	AGE
Same address	1	>=1	-
Different address	2	0	-
Child under 1	3	-	0

SEG20<-SEGE

	Standard Key Variable	GHS Variable
Value Label	SEG20	SEGE
Employers and Managers, large businesses	1	1,2
Employers small businesses	2	3
Mangers small businesses	3	4
Professional self-employed	4	5
Professional employed	5	6
Ancillary workers and artists	6	7
Foreman and supervisors - non-manual	7	8
Junior non-manual	8	9
personal services	9	10
Foremen and supervisors -manual	10	11
Skilled manual	11	12
Semi-skilled manual	12	13
Unskilled manual	13	14
Own account non-professional	14	15
Farmers employers and Managers	15	16
Farmers - own account	16	17
Agricultural workers	17	18
Armed forces	18	19
Inadequately stated	19	-8
Children and Never worked, F-T students	20	-9, -6, 20

TENURE6<-TENURE

	Standard Key Variable	GHS Variable
Value Label	TENURE6	TENURE
Owner occupier mortgaged	1	3
Owner occupier outright	2	2, 1
Rented with job or business	3	4,5

rented from LA	4	6,7
rented privately furnished	5	10
rented privately unfurnished	6	8,9,11

CARS4<-NCARS1

	Standard Key Variable	GHS Variable
Value Label	CARS4	ncars1
0	0	1
1	1	2
2	2	3
3+	3	4

LTILLHH2

	Standard Key Variable	GHS Variable
Value Label	LTILLHH2	
> 0 persons in household with long-term limiting illness.	1	DERIVED
No persons in household with long-term limiting illness.	2	DERIVED

DEPCHILD2

	Standard Key Variable	GHS Variable
Value Label	DEPCHILD2	
>0 Dependant children in household	1	DERIVED
No dependant children in household	2	DERIVED

PENSIONHH2

	Standard Key Variable	GHS Variable
Value Label	PENSIONHH2	
>0 persons of pensionable age in household	1	DERIVED
No persons of pensionable age in household	2	DERIVED

RESIDENTS 4

	Standard Key Variable	GHS Variable

Value Label	RESIDENTS4	
0	0	DERIVED
1	1	DERIVED
2-5	2	DERIVED
6+	3	DERIVED

4) Procedures for dealing with non-standard and missing responses.

There are 412 cases affected by non standard and missing responses. Where possible, these values are imputed from other variables. If it is not possible to do this then in most cases a wild-card value is coded. Wild-cards will match any value but at 0.5 match value.

AGE94<-AGE

No non-standard or missing responses.

COBIRTH22<-COB

56 cases of non-responses(99) coded as wild-card (-1)
1 case of 'answer too general' (23) coded as wild-card(-1)

PRIMECON11<-ECSTILO+WKSTATE+SEGE

ecstilo: 136 cases of non-response coded using ecstaa follows:

if ecstaa = 5 then PRIMECON11 is coded as unemployed (6)
if ecstaa = 8 then PRIMECON11 is coded as retired (9)
if ecstaa = 10 then PRIMECON11 is coded as student (7)
if ecstaa = 9 then PRIMECON11 is coded as other inactive (10)
if ecstaa = 11 then PRIMECON11 is coded as other inactive (10)
otherwise PRIMECON11 is coded as wild-card (-1)

QUALLEVEL4<-EDLEV

16 cases of non-response coded using TEA as follows:
If tea<20 then code QUALEVEL4 no higher qualifications (0)
otherwise code as wild-card (-1)

FAMTYPE4<-FUT20

No non-standard or missing responses.

LTILL2<-LIMILL

59 cases of non response coded using LSILL as follows:

If lsill = 2 then LTILL2 is coded as no long term limiting illness(2)
If lsill = -8 then LTILL2 is coded as wild-card(-1)
If lsill = 1 then LTILL2 is coded using RESTACT as follows:
If restact = 2 then LTILL2 is coded as no long term limiting illness(2)
If restact = -8 then LTILL2 is coded as wild-card(-1)
If restact = 1 then LTILL2 is coded as has long term limiting illness(1)

MARCONA5<-MARSTAT

No non-standard or missing responses.

ETHNIC10<-ORIGIN

69 cases of non-response(-8) coded as wild-card(-1)

MIGRANCY3<-RESLEN5+AGE

44 cases of non-response(-8) coded as wild-card(-1)

SEG21<-SEGE

18 cases of non-response (-8) coded as inadequately stated()

TENURE6<-TENURE

31 cases of caravan (13) removed from sample.
6 cases of non-response(12) coded as wild-card(-1).

CARS4<-NCARS1

14 cases of non-responses coded as wild-card(-1)

LTILLHH2

Where LTILL2 for an individual has a wild-card coding(-1) LTILLHH is coded as wild-card unless the LTILL2 coding for another individual within the same household is coded as 1 in which case LTILLHH2 is coded as 1.

DEPCHILD2

No non-standard or missing data (age94).

PENSIONHH2

No non-standard or missing data (age94/sex2).

RESIDENTS 4

No non-standard or missing data.

5) Relationship between standard and GHS geographies.

REGION12<-REGION

REGION12 is the standard geography used for matching between the GHS and the 1% household SAR. It is a direct translation.

	Standard Geography	**GHS Variable**
Value Label	REGION12	REGION
North	1	01, 02
Yorks. and Humber	2	03, 04
East Midlands	3	07
East Anglia	4	10
Inner London	5	11
Outer London	6	12
Rest of South East	7	13, 14
South West	8	15
West Midlands	9	08
North West	10	05, 06
Wales	11	16, 17
Scotland	12	18-22

REGION20<-REGION

REGION20 is the standard geography used for the matching between the GHS and the 2% individual SAR. It is not possible to use an direct equivalent of the GHS region variable because two SAR areas (254: Berwickshire, East Lothian etc. 129: Castle point Maldon Rochford) fall cross region boundaries.

	Standard Key Variable	**GHS Variable**
Value Label	REGION20	REGION
North Metropolitan	1	01
North non-metropolitan	2	02
Yorks. and Humberside Met	3	03
Yorks. and Humberside Non-met	4	04
North West Met	5	05
North West Non-Met	6	06
East Midlands	7	07
West Midlands Met	8	08
West Midlands Non-Met	9	09
East Anglia	10	10
Inner London	11	11
Outer London	12	12
Rest of South East	13	13,14
South West	14	15
Wales Northern	15	16
Wales Southern	16	17
Northern Scotland	17	18
Glasgow	18	20
Strathclyde	19	21
South & Central Scotland	20	19, 22