Adding Area-based Classifications to the Samples of Anonymised Records (SAR) from the 1991 Census

Angela Dale and Stan Openshaw

1. Background

In 1996 ONS (then OPCS) agreed in principle that an area-based classification could be added to both of the SAR data sets. It was evident that the addition of an area classification would greatly enhance the value of the SARs as a research resource. This would entail adding a ward or ED-level classification code to each household in the Household SAR and to each individual in the Individual SAR. These codes would thereby provide an element of the small area information which is an essential feature of the census and which is perceived by many to be a shortcoming of the SARs. It would also open up possible new areas of research related to the SAR data; particularly, multilevel modelling and ecological-aggregational inference error studies. It would also provide a basis for evaluating the properties of area based classifications - an area of contemporary research and practical interest.

The addition of an area-level classification to the SARs was seen to be of considerable academic value - for example, greatly enhancing the scope for modelling multilevel area effects. In analysing unemployment, the addition of an area-level classification would provide a much better indicator of the locality than can currently be obtained from knowledge of either the SAR area (large LA) in the 2% sample or the region in the 1% sample. Similarly, in the analysis of long-term limiting illness, it is important to be able to include in a multilevel model characteristics of an area at a much finer level of geography than is currently possible. An area-level classification provides the means for doing this whilst protecting information about the actual ED of residence. It would also enhance the marketability of the SARs to the commercial sector and to health authorities, both of whom make considerable use of such classifications.

Subsequently, the ESRC's Research Resources Board made available funding to pay for the cost of adding two sets of area classification codes to the SARs. The purpose of this short paper is to explain why, how, and which classifications were added to the SARs.

2. Choice of area level classifications

A number of different commercial and academic small area-level classifications are available, almost all developed using Small Area Statistics from the Census, and sometimes including other data sources such as local unemployment statistics. Most are ED or postcode based and provide a way of attaching a composite indicator of the local area to an individual or household. Because of the very large number of EDs in the country (about 150,000), each category of an area-level classification typically applies to several thousand EDs - although this will vary with the number of categories in the classification.

ONS were only willing to allow one classification to be added to each SAR and then only if special precautions were taken to avoid the risk of any possible confidentiality problems. After consulting fairly widely across all sectors of SAR users it was concluded that there we should ask ONS to add their own classification to the 1% Household SAR and add GB Profiles (or some variant of it) to the 2% SAR.

The ONS classification is attractive because there is considerable detail on the method of construction (Wallace, Charlton and Denham, 1995) and it was regarded by many of those

consulted as of high quality. It provides continuity with district level classifications produced for earlier censuses. It represents the 'official' classification which may well be widely used by local authorities and health authorities. However, it is a poor spatial resolution product (being restricted to a ward scale of geography) and it is noted that all the commercial sector and research applications involving small area classifications are at a far finer geographic scale; mainly unit postcodes or census EDs (Sleight, 1997).

GB Profiles was developed in a research project funded by an ESRC Research Grant (Openshaw, Birkin, and Blake, 1994-5) that was separate from the ESRC 1991 Census Programme. It is well regarded by many academics although it is less well known than some of the commercial sector classifications. However, it is currently the only geodemographic system freely available for teaching and academic research purposes in the UK. Access to the commercial sector equivalents has not yet been organised in a systematic manner and the precise details of their construction tend to be regarded as a commercial secret. Additionally, it is likely that some users of the enhanced SAR data may be interested in developing critiques of geodemographics and the use of a commercially available product raised the prospect of litigation if such criticism was seen as reducing sales of a commercial product. These problems could be avoided if GB Profiles is used. There is detailed information on the derivation of the GB Profiles classification (Openshaw, Blake and Wymer, 1994) and this is also available on the Web. Also the computer codes used have been published (Openshaw, 1994) and described in some detail (Openshaw and Wymer, 1996). It is believed that this classification is at least as good as the commercial sector equivalents that were based on 1991 census data as it probably used equivalent or superior technology in its construction. A further key consideration is that Professor Openshaw offered to optimise the structure of the classification to provide the finest resolution consistent with meeting ONS's confidentiality requirements.

It was therefore decided that the ward-level ONS classification should be added to the Household SAR and that GB Profiles or some modified derivative should be added to the 2% Individual SAR. Each is described in the following sections.

3 The ONS ward classification

The SAR variable ONSCLASS was created from the ONS ward classification. A value has been attached to every individual in the Household SAR and provides a descriptor of the ward in which the household is located. Individuals in the same household share the same ONSCLASS value. However, because of ONS confidentiality requirements some amendments have been made to the standard classification to reduce possible disclosure risks.

3.1 Description of clusters

The ONS ward classification assigns wards in England and Wales and postcode sectors in Scotland to one of 14 Groups and 43 Clusters according to their characteristics based on 1991 Census data. The derivation of the classification is described in Wallace, Charlton and Denham (1995) and, more fully, in Wallace and Denham (1996). The 14 Groups, which form the basis of the SAR variable ONSCLASS, are described in Table 1.

3.2 Amendments to avoid disclosure risks

In producing the variable ONSCLASS, a threshold was used which required that, if a Group classification code was represented in any SAR region, it would apply to at least 5 wards in that region. Where there were fewer than 5 wards, the Group code in the standard ONS classification was combined with a another, similar, Group code in the same region. This

change was necessary because certain SAR records could be identified as having been enumerated in a specific ward; e.g. a SAR record with a Group 1 code in Inner London could only come from a single ward, which could be identified by reference to a full listing of the ONS classification.

The ONSCLASS variable therefore consists of 27 codes, not all of which appear in every region. Codes 1-14 are the same as the standard ONS classification whilst ONSCLASS codes 15-27 are formed by combining different Groups from the ONS classification. For example, a user will know that an East Midlands SAR record with ONSCLASS code 17 will be either from one of 104 Group 6 wards or from one of 4 Group 10 wards. This amended classification is shown in Table 2. This may appear both confusing and complex but an attempt has been made to combine groups in a meaningful way. For example, groups characterising areas of deprivation were combined with similar groups, rather than with groups which normally apply to more prosperous areas. However, different combinations were necessary in different regions. For example, in the North and East Midlands regions, Group 10 wards have been combined with Group 6 wards to form ONSCLASS code 17. However, in the West Midlands, Group 10 wards have been combined with Group 4 wards to give ONSCLASS code 26. Finally, the ONS classification was not produced for certain wards and postcode sectors with very low populations. Where SAR cases are from these areas, the predominant classification code from neighbouring areas has been applied.

It is hoped that these cluster codes will be useful as contextual proxy descriptors of the local area within which the SAR households are located. This might be regarded as being most useful for multilevel modelling applications and, perhaps, less useful in trying to identify ecological fallacies in an area based census classification. The latter application is met by the GB Profile codes added to the 2% individual SAR.

4 The GB Profiles ED level classification

4.1 Background

The GB Profiles ED level census classification is described in Openshaw et al (1995). It can be downloaded on to a PC or used at MIDAS. The idea was to offer multiple different census classifications with varying numbers of clusters in them. In practice only a small number were ever made public because of historic (but now irrelevant) constraints on PC hardware (in the mid 1990s when the research was performed). The published and non-published GB Profile Classifications were offered as a basis for adding one on to the SAR. The ONS modified ward level classification provided a meso scale contextual variable so the objective now was to identify an acceptable mechanism for adding ED level cluster codes on the 2% person SAR. For this exercise the ONS rule of thumb confidentiality criteria was that when a classification for 145,716 EDs into M clusters was crosstabulated by SAR geography (with 278 areas) that all the cell counts were either zero or greater than 10. As with the ward classification the problem was what to do with clusters in the small area classification which failed the ONS confidentiality constraint.

4.2 Amendment strategies

The options were as follows.

1. Flag these eds as missing. (This is the simplest solution but maybe a large chunk of the SAR data would be affected and it was likely that some types of area and certain regions of

the country would be more affected than others, thereby reducing the utility of the small area codes. Table 3 shows that, depending on how many clusters were used, a large part of the SARs would be affected. The purpose of the exercise was to add a good small area classification and this implied that a reasonable number of clusters should be used (circa 40 to 80) rather than a few (circa 5-15), otherwise little benefit would be gained.

2. Instead of flagging failed eds as missing they could be assigned the cluster code of the nearest ed that was not flagged as missing. However, ONS were keen that no distinction could be made between real cluster codes and nearby ones. Given the large numbers of affected eds in Table 3 this was clearly unsatisfactory since the nearest neighbour assignments would be of variable accuracy and representativeness. It would certainly devalue the classification as no distinction is made between real clusters and nearby ones and maybe lead others to false conclusions about the quality and properties of GB Profiles. This was unacceptable.
3. Repeat the modifications process used on the ONS ward classification. However, whilst better than (2) it was still non-ideal. It was judged to be far too complex and was considered inappropriate given the aim of using GB Profiles to provide a viable small area classification code. Also there was access to the GB Profile classification methodology so it might be better to re-do the classification (rather than merely re-aggregate or juggle the cluster codes) in order to meet the confidentiality criteria.
The fourth alternative was to identify the affected eds and then assign them to their second or third nearest etc cluster in the classification which would meet the confidentiality criteria. These would then be flagged as being re-assigned by a negative cluster code. This would provide option (1) as a by-product but at the same time try and preserve as much as possible of the quality of the original classification. It is noted that some re-assignments would probably have little noticeable impact on the quality of the classification because of fuzzyness and redundancy in all multivariate classifications. For instance, the allocation of an ed to cluster 17 might be extremely marginal (i.e. a small difference in a real number) and often 27 or 42 or 13 might almost be just as good. So in principle this strategy seemed better than the alternatives and ONS agreed to accept this option provided that there were either no or multiple re-assignments in each SAR area to prevent an attempt at reverse engineering.

4.3 Amendments algorithm

Strategy 4 was adopted although it involved a degree of extra work and effort. The problem now was how best to optimise the re-assignment process so that as few eds as possible were affected. It is possible that some eds would be saved because of the re-assignment of others if the process could be made gradual rather than occur as a single step one. After some experimentation the following algorithm was devised and ONS confirmed that it produced acceptable results.

Step 1 select desired number of clusters
Step 2 classify all 145,716 census eds using a Kohonen self-organising map classifier and keep the neuronal weights
Step 3 set K=1
Step 4 crosstabulate the ed classification clusters by SAR areas and flag as unclassified any eds for which the cell count is greater than zero and less than or equal to K
Step 5 re-assign these unclassified eds to the “nearest” valid cluster
Step 6 repeat Step 4 to 5 until no unclassified eds are left
Step 7 set K=K+1 and if K is less than or equal to 10 return to Step 4.
Step 8 check that there have been multiple re-assignments to any cluster used in re-assignment

process

The remaining questions concerned the desired number of clusters and whether the modified classification still looked anything like the original. The first question was resolved by experimenting with various numbers of clusters. The eventual solution was 49 corresponding to a Kohonen self-organising two dimensional map neuron array with 7 rows and 7 columns. Note that the use of a self-organising map (rather than K-means) reflected the belief that the topology preserving properties of this classifier would be beneficial when it came to the re-assignment process. The answer to the second question can be summarised as "yes", the labelling and interpretation of both the original and the modified classifications were very similar.

4.4 Interpretation of the SAR GB Profile 49 cluster system

The 49 cluster GB profiles classification is based on 80 census variables, see Appendix A. These were classified using the self-organising Kohonen neural net. The 49 clusters have been labelled and the thumbnail descriptions are given in Table 4. These are, of course, like all cluster labels subjective generalisations and the diagnostics on which they are based will be made available so that users can "make-up" their own area profile descriptions.

Finally, the full postcode and ED directories for this classification will be accessible from MIDAS via the GB Profiles system.

5 Conclusions

The paper summarises the process by which two different area classification codes were added to the SAR data. It documents a unique three-way collaboration between researchers at three different institutions (Manchester and Leeds Universities and ONS). The hope is that this unique enhancement of microcensus data will create additional research opportunities and add value to the ESRC's investment in the SAR data sets.

Reference

Wallace, M., Charlton, J. and Denham, C. (1995) ‘The new OPCS area classification’, *Population Trends*, 79, 15-30

Wallace, M. and Denham, C. (1996) The ONS classification of wards. SMPS 60. (London: HMSO)

Table 1

Group	Description
1	Suburbia
2	Rural Areas
3	Rural Areas with mixed economies
4	Industrial & Manufacturing Towns
5	Middling England
6	Prosperous wards
7	Purpose-built, Inner City estates
8	Established Owner-Occupiers
9	Armed Forces bases
10	Metropolitan professionals
11	Deprived City Areas
12	Lower Status Owner Occupiers
13	Mature Populations
14	Deprived Industrial Areas

Table 2 A Derivation of ONSCLASS codes

Region	ONSCLASS	*formed by combining*	ONS ward code	(Number of wards)	*and*	ONS ward code	Number of Wards
East Anglia	**15**		**4**	33		**14**	3
	16		**12**	14		**11**	1
East Midlands	**16**		**12**	37		**11**	1
	17		**6**	104		**10**	4
	18		**14**	20		**7**	2
Inner London	**19**		**12**	6		**1**	1
	20		**7**	80		**4**	2
	21		**11**	134		**14**	4
North West	**22**		**3**	74		**9**	1
North	**23**		**2**	62		**9**	2
	17		**6**	32		**10**	4
	16		**12**	64		**11**	4
Outer London	**24**		**8**	63		**13**	3
Rest of South East	**18**		**14**	4		**7**	1
	16		**12**	61		**11**	3
Scotland	**25**		**4**	188		**12**	1
South West	**16**		**12**	49		**11**	1
	18		**14**	5		**7**	2
Wales	**18**		**14**	39		**7**	1
	23		**2**	153		**9**	4
West Midlands	**18**		**14**	30		**7**	1
	26		**4**	73		**10**	4
Yorkshire & Humberside	**27**		**7**	4		**11**	3

Table 3 Lost eds

Number of Clusters	Number of EDs affected	Worst area lost % of eds
2	0	0
5	630	7
10	3111	10
15	4762	15
20	6030	20
25	9003	25
30	10937	33
35	11855	32
40	14151	30
45	17472	38
50	19213	51
55	21632	53
60	23876	52
65	24853	52
70	26831	53
75	27341	58
80	30351	60
85	32970	63
90	34877	70
95	37107	74
100	39542	75
110	43068	82
120	46548	85

Table 4. GB Profiles 49 Cluster Labels

Cluster 1: High LLTI, retired pensioners, council housing, no car
Cluster 2: Outright-owners, detached housing
Cluster 3: Semi-detached, privately owned with mortgages, high car ownership, families,professional jobs
Cluster 4: Elderly, retired home-owners
Cluster 5: Asian, high unemployment, overcrowded, terraced housing
Cluster 6: Small, semi-detached council housing
Cluster 7: Terraced housing, Council or Housing Association, couples without children
Cluster 8: White owner-occupiers with cars, mixed types of household
Cluster 9: Retired couples, white, outright owners in semis
Cluster 10: Couples without children, detached houses, owned outright, 2+ cars
Cluster 11: White, large detached houses, owner-occupiers, with cars in professional jobs
Cluster 12: Young couples without children, employed with mortgage, recent movers
Cluster 13: Indian(+ other ethnic minorities) in rented/terrace housing
Cluster 14: Mid-life couples with children and mortgage, in work with cars, detached house
Cluster 15: Terraced housing, white working class
Cluster 16: Rural/farming community
Cluster 17: Elderly, retired, flats in social housing, no cars
Cluster 18: Black and Asian household, singles/lone parents, flats in social housing, high unemployment, no cars
Cluster 19: Private rented bedsits, shared facilities, young, single and mobile
Cluster 20: Armed forces, young families with children, recent movers
Cluster 21: Singles/lone parents, social housing, public transport to work, high LLI
Cluster 22: Black, Chinese, Indian, terraced housing, public transport to work
Cluster 23: Terraced housing, LA/HA rented, no central heating, no car
Cluster 24: Couples with children in LA housing, terrace, singles, no car
Cluster 25: Couples with dependent children in LA housing, terraced, manual workers
Cluster 26: Elderly retired, single in LA housing, flats and couples with no children
Cluster 27: Singles/lone parents, LA/HA housing, flats, unemployed
Cluster 28: Young employed couples without children, flats, recent movers
Cluster 29: Minority ethnic groups in owner-occupation or private renting, semi-detached, 2+families
Cluster 30: Young couples without children, terraced housing, no heating, no cars
Cluster 31: White, middle-ages couples with children, home-owners large, detached houses, 2+ cars, professional
Cluster 32: Couple with children in owner-occupation, semi-detached housing, mortgage,economically active
Cluster 33: Chinese and black groups in LA housing/blacks in private housing, overcrowding, flats,no cars, public transport to work
Cluster 34: Rural/farming community - self employed/farmers
Cluster 35: Couples with dependent children in LA housing; semi-detached housing, no car
Cluster 36: Semi-detached housing, owner-occupied; all else average
Cluster 37: Area with no distinctive features(semi-det, manual workers)
Cluster 38: Large house, detached, professional workers with higher qualifications, 2+ cars, couples with children, students
Cluster 39: Single people in owner-occupied housing, terraced
Cluster 40: Terraced housing, owner-occupied with mortgage, working, car, not singles.

Cluster 41: Older couples, no children, retired, detached, owned outright
Cluster 42: Professionals/banking, finance, higher quals, owner-occupiers
Cluster 43: Rented LA/HA, semi-detached, no heating
Cluster 44: Rented LA/HA housing, not detached, not 2+ cars
Cluster 45: Lone parents, flats, rented from LA/HA, high unemployment
Cluster 46: White, detached, 2+ cars, families, drives to work
Cluster 47: Single, owner-occupiers/Chinese private renting, mobile population, flats, higher qualifications, working in Banking and Finance
Cluster 48: White, car to work
Cluster 49: LA rented, terraced and semis, manual workers

Appendix 1. List of Variables used in creating the GB Profiles classification

Variable	1	Aged 0 - 4
Variable	2	Aged 5 - 14
Variable	3	Aged 15 - 24
Variable	4	Aged 25 - 44
Variable	5	Aged 45 - 64
Variable	6	Aged 65 - 74
Variable	7	Aged 07 75 - 84
Variable	8	Aged 85+
Variable	9	Total married population
Variable	10	Single population
Variable	11	Total retired (pensioners)
Variable	12	Working Women (excluding Govt Sch.) (S08)
Variable	13	Total 'Lone' Parents
Variable	14	Students (16+) in term-time addresses
Variable	15	White
Variable	16	Black
Variable	17	Indian
Variable	18	Pakistani
Variable	19	Bangladeshi
Variable	20	Chinese + Other
Variable	21	Black (grps) and Owner & privately rented
Variable	22	Indian, Pakistani, Bangladesh and Owner & privately rented
Variable	23	Chinese & others and Owner & privately rented
Variable	24	Black (grps) and council rented
Variable	25	Indian, Pakistani, Bangladesh and council rented
Variable	26	Chinese & others and council rented
Variable	27	Movers last year
Variable	28	Pensioner migrants
Variable	29	Owned Outright
Variable	30	Mortgaged
Variable	31	Private Rented
Variable	32	Rented from Housing Authority, Local Authority, New Towns
Variable	33	Detached
Variable	34	Semi-detached
Variable	35	Terraced
Variable	36	Flats
Variable	37	Bed-sits
Variable	38	No central heating
Variable	39	Lacking bath and shower
Variable	40	Few cars
Variable	41	2+ cars
Variable	42	Households with more than 1.5 persons per room
Variable	43	Number of Households with 7+ rooms
Variable	44	Couple Households, aged 16-24 without child(ren)
Variable	45	Couple Households, aged 16-24 with child(ren)

Variable	46	Couple Households, aged 25-34 without child(ren)
Variable	47	Couple Households, aged 25-34 with child(ren)
Variable	48	Couple Households, aged 35-54 without child(ren)
Variable	49	Couple Households, aged 35-54 with child(ren)
Variable	50	Couple Households, aged 55-75 plus
Variable	51	No Family Households / Owner
Variable	52	No Family Households / Council
Variable	53	Married + cohabiting Couple no children / Owner
Variable	54	Married + cohabiting Couple no children / Council
Variable	55	Married + cohabiting Couple + dependent children / Owner
Variable	56	Married + cohabiting Couple + dependent children / Council
Variable	57	2+ Family Households / Owner
Variable	58	2+ Family Households / Council
Variable	59	Households with dependants
Variable	60	Economically active residents 16+
Variable	61	self-employed
Variable	62	Total Economically Active. unemployed
Variable	63	Agriculture/Forestry/Fishing
Variable	64	Energy, Water & Mining
Variable	65	Manufacturing
Variable	66	Construction
Variable	67	Distribution & Catering
Variable	68	Transport
Variable	69	Banking & Finance
Variable	70	Professional (1,2,3,4)
Variable	71	Intermediate & Junior Non-manual (5,6)
Variable	72	Manual (8,9,12; 7,10; 11)
Variable	73	Agricultural (13, 14, 15)
Variable	74	Armed Forces (16)
Variable	75	Workers with higher degrees
Variable	76	Workers with other qualifications
Variable	77	Total persons with Long Term Limiting Illness (S12)
Variable	78	Train & Bus
Variable	79	Car
Variable	80	Work at home