CAN ADMINISTRATIVE DATA BE USED TO CREATE A GEODEMOGRAPHIC CLASSIFICATION?

Mildred Oiza Ajebon [1] and Paul Norman [2]

1 Department of Geography, University of Durham, Durham, DH1 3LE, UK
Email: m.o.ajebon@durham.ac.uk
Tel: +447405890795

2 Centre for Spatial Analysis & Policy, School of Geography, University of Leeds, Leeds, LS2 9JT, UK
Email: p.d.norman@leeds.ac.uk
Tel:+44 (0)113 34 38199 Fax:+44 (0)113 34 33308

Acknowledgements

This work used Census data obtained via MIMAS' CASWEB and GIS boundary data obtained via EDINA's UKBORDERS; services supported by ESRC and JISC. These data are Crown copyright and are reproduced with permission of OPSI. Adapted data from the Office for National Statistics and obtained via the Neighbourhood Statistics and NOMIS websites licensed under the Open Government Licence v.2.0 are also used. All data are subject to Crown Copyright.

Abstract

This paper aims to contribute to the wider research scheme of the ONS 'Beyond 2011' project by assessing the feasibility of creating geodemographic classifications from administrative statistics as a way of eliminating the need for a full population survey. The classification is created using K-Means clustering algorithm which is then compared with OAC super-groups as a benchmark in maps and cross-tabulations. Results show similar classification of area types and health variations in England suggesting that the range of administrative datasets examined in this study could be explored as viable alternatives to the traditional census approach.

Keywords: Geodemographic Classification, Health Variation, Administrative Data

1.0 Introduction

The decennial census, though a complete source of socioeconomic data on UK population, has been criticized as being too costly and becoming increasingly difficult to carry out due to contemporary changes in the way society is organised. The challenges of high population mobility, opportunities created by advances in information technology which has increased the efficiency in the way data on several aspects of the population is stored, and the need for more timely and up-to-date delivery of demographic data across the UK, all seem to suggest alternative ways of collecting, organising and disseminating detailed and up-to-date population data ONS (2011). Hence, the UK Statistical Authority (UKSA) "Beyond 2011" programme was set up to examine the feasibility of replacing the traditional census approach with administrative data already being held on the population. A logical starting point for the programme would be to search for similarities in patterns identified by both administrative and census datasets on a wide range of topics. This study therefore, aims to assess the potential of creating area classification models from available administrative data sources with the 2001 OAC Super-Groups as a benchmark. The geodemographic model is chosen for a pilot study because it is one of the most widely used socioeconomic models created from the decennial census for public sector planning and business targeting. The reliability of the classification is tested using independent data sets not included in classification.

The choice of testing the classification against selected census ill-health indicators is informed by the evidence in literature on the use of geodemographics to explain health inequalities health across England, The work of Dedman et al. (2006) demonstrated that geodemographic systems can be used to classify areas according to health needs by clearly showing where high and low illness rates might be expected. Shelton et al. (2006) developed a geodemographic characterisation of mortality patterns in England. Using sex and age standardised mortality data for nine causes of death, he calculated SIRs and found patterns of mortality to reflect socioeconomic circumstances with the more deprived areas suffering poorer health outcomes. Petersen (2009) also found that health inequality can be illustrated based on social area types. Thus, efforts have been made to show that strong relationships exist between population health and area types. However, no attempts at comparing geodemographic classifications created from administrative data sources as potential replacements of the census-based area classification were uncovered. This study represents one of the first efforts directed at filling this research gap. It examines the feasibility of constructing a geodemographic classification for small areas in England from administrative data sources. It examines how the classification compares with the census-based Super groups and whether the new classification can be used to predict geographical patterns of inequalities in health across England.

2.0 Data and Methods

The primary scale of analysis chosen for this study are the 32,482 Lower Super Output Areas (LSOAs) of England being the geographies for which small area administrative statistics are published regularly to enable analysis of patterns over time (Neighbourhood-Statistics, 2004). The GIS boundary data for LSOAs are obtained from the United Kingdom Baseline Reference Database for Education and Research Study (UKBORDERS) available at EDINA (2012). It is worthy of note that the choice of variables for this study is highly limited by the availability of administrative statistics. All the datasets are derived from 100% administrative data sources and are a product of the National

Statistics. They are produced to a high statistical standard and accuracy. All datasets found to be negatively skewed were log-transformed (***LN(Data +1))*** to near normal distributions which has been found to be well adapted to socioeconomic count data (Rogerson, 2010). With the exception of council tax band as a proxy measure of housing, the strength of relationships between equivalent pairs of variables are shown to be strong in table 1. The classification of LSOAs into six socioeconomic groups as measured by available administrative data is created using the functionality of the SPSS k-means iterative clustering algorithm. Please refer to (Birkin and Clarke, 1998, Birkin and Clarke, 2009, Vickers and Rees, 2006, Vickers and Rees, 2007) for comprehensive details on creating geodemographic classification of areas. The alternative classification created from administrative data sources in this study is generally labelled Geo-Social Area Classification (GSAC). The profile names and pen portraits are derived from most dominant variables in each cluster.

	JSAC	Lone Parent	Council Tax Band	Incapacity Benefit	Pension Claimants	Census Unemployment	Census Lone Parents	Census Rented	Census Pensioners	Census LLTI
JSAC	1									
Lone Parent	.674	1								
Council Tax Band	.537	.469	1							
Incapacity Benefit	.726	.641	.682	1						
Pension Claimants	-.127	-.079	.119	.104	1					
Census Unemployment	**.865**	.758	.551	.755	-.055	1				
Census Lone Parents	.649	**.835**	.555	.657	-.130	.709	1			
Census Rented	.614	.645	**.512**	.557	.067	.680	.658	1		
Census Pensioners	-.187	-.112	.057	.054	**.970**	-.093	-.194	-.025	1	
Census LLTI	.453	.435	.578	**.746**	.614	.543	.385	.424	.589	1

Table 1: Correlation matrix of the main census and administrative variables

The correlations between equivalent pairs of deprivation variables are highlighted in beige. All correlations are significant at p=0.00. Incapacity Benefit (IB), Limiting Long Term Illness (LLTI); Job Seekers Allowance Claimant (JSAC);

3.0 Analysis

3.1 Administrative data-based Area Classification of England

The classification created from administrative data is generally named ‘Geo-Social Area Classification (GSAC)’. Tables 2 shows profile labels and pen portraits of the six area types labelled after the dominant variables shaping the social character of the clusters. The map of the six clusters and the ONS Supergroups is shown in Figure 1.

Social profiles			
Area Type	**Dominant Characteristics**	**Area**	**Dominant Characteristics**
1: Struggling Families	Persons paying low council tax (bands B & C) Pension Claimants Aged 50 and over Paid care givers Incapacity benefit claimants	4: Suburbia	Resident population mainly elderly aged 60 and over Medium to high council tax payers (Bands D – G) Pension claimants
2: Typical Urban Living	Middle-aged resident population mainly older adults (aged 25-49) Paying council tax D- E High Population Density Lone Parents Job seekers	5: Outer Urban	High council tax payers Residents mainly aged 50-59
3: Deprived Communities	Persons paying low council tax (bands A) High children population (0-15) Job seekers Lone parents Incapacity benefit claimants Paid carers	6: Young Urban Families	Young resident population aged 16-24 High population density Low council tax (Band A)

Table 2: Cluster labels and the variables defining the socio-economic characteristics of LSOAs in England

A visual comparison of the maps in Figure 2 shows some spatial similarities between the ONS Super-Groups and the alternative GSAC. This inequality is well defined by both classifications in regions like the North-East, South Yorkshire, North-West and the Midlands. The ONS ‘Country-side’ and the GSAC ‘Suburbia’ which contain affluent LSOAs and the elderly population reflect more suburban distributions. The ethnic dimension is largely missed in the GSAC classification of urban areas due to the lack of readily available small area administrative data on ethnicity for inclusion in the area classifications at the time of this study. Overall, the GSAC appears to identify areas of socioeconomic disadvantage more distinctively. This is as expected since the main administrative variables available for inclusion in the classification relate to various types of economic deprivations. The ONS classification demonstrates a smoother pattern of socio-economic structuring of the population. This pattern is examined statistically by relating the classifications to independent observations to examine how they perform in reflecting socio-economic stratification.

3.2 Cross Tabulation-Based Comparisons of Geodemographic Classifications

Tables 3a and 3b show the results of the cross tabulations of the ONS Super-Groups and the GSAC clusters which is a widely used objective method of determining the ecological equivalence of area classifications (Voas and Williamson, 2001, Webber and Butler, 2007). Each cell on the table shows the proportion of the total population share of LSOAs common to the subclasses of both classifications. The suburbia neighborhoods is classified as an equivalent of the Super-Group country side and urban fringe area types. The GSAC deprived communities are found to be similar to the ONS multicultural and disadvantaged groups. These clusters and most urban areas appear under-represented in the GSAC. Table 3b contains the index scores which quantify the degree of appropriateness of these proportions. An index score of 0 shows lack of representation and absence of the target cluster in the benchmark classification. 50 means that the target cluster is half represented as expected, 100 depicts equal representation on both classifications, an index value greater than 100 indicate an over-representation and 200 means that the target cluster is twice as represented in that order (Boyle et al., 2004). The indices shown in Table 3b indicate higher ecological correspondence between the urban clusters of both classifications. The GSAC appears to more clearly, distinguish areas of socio-economic disadvantage than the ONS Supergroups.

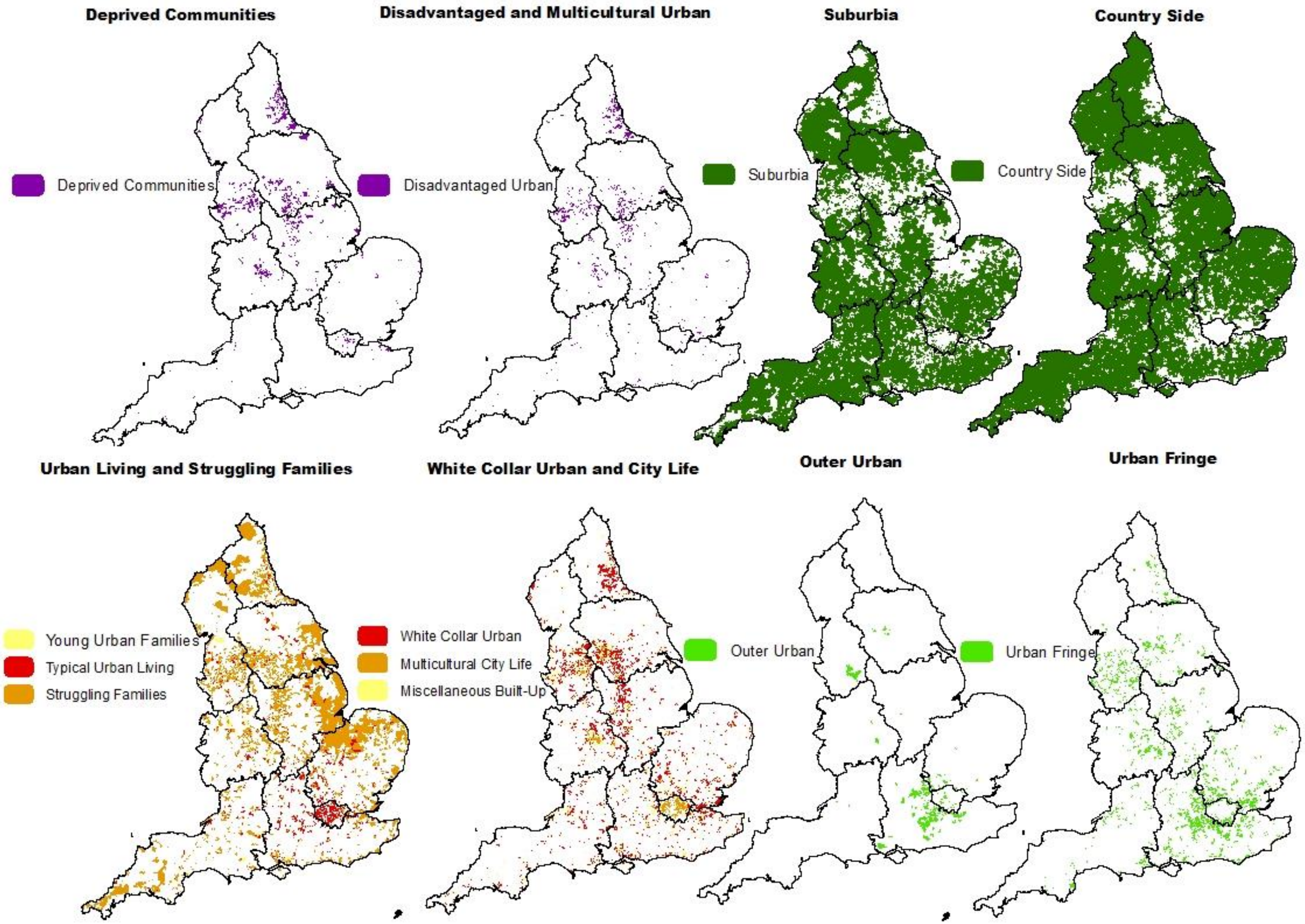


Figure 1: Clusters of ONS LSOA-level geodemographic classification Super-Groups compared with the alternative 'Geo-Social Area Classification' (GSAC)

A	COUNTS OF LSOAS	ONS SUPERGROUP TYPES FOR LSOAs							Total Number of LSOAs	% LSOAs
		1	2	3	4	5	6	7		
		Country Side	Professional City Life	Urban Fringe	White Collar	Multicultural City Life	Disadvantaged Community	Miscellaneous Built-Up		
GEO-SOCIAL AREA CLASSIFICATION (GSAC)	1: Struggling Families	780	212	609	3840	259	1063	3802	10565	32.5
	2: Typical Urban Living	65	1411	959	873	1789	98	850	6045	18.6
	3: Deprived	11	53	0	86	1457	3236	898	5741	17.7
	4: Suburbia	3120	387	3449	1440	73	1	536	9006	27.7
	5: Outer Urban	59	202	308	0	0	0	1	570	1.8
	6: Young Urban Families	16	400	28	7	47	0	57	555	1.7
Total Number of LSOAs		**4051**	**2665**	**5353**	**6246**	**3625**	**4398**	**6144**	32482	100.0
% LSOAs		**12.5**	**8.2**	**16.5**	**19.2**	**11.2**	**13.5**	**18.9**	100.0	

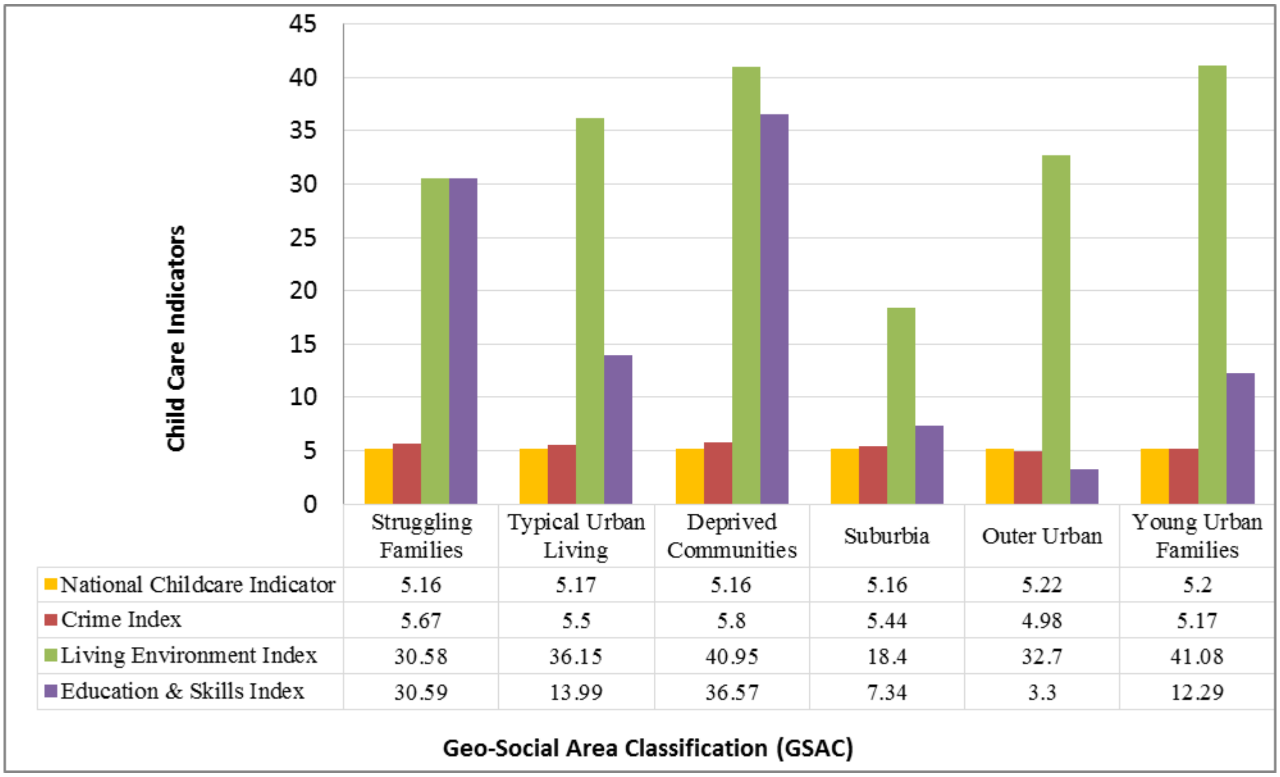
B INDICES		1	2	3	4	5	6	7
	Population Share	Country Side	Professional City Life	Urban Fringe	White Collar	Multicultural City Life	Disadvantaged Community	Miscellaneous Built-Up
1: Struggling Families	0.325257	59 (10.0)	24 (4.1)	35 (5.9)	189 (32.8)	22 (3.7)	74 (12.5)	190 (32.0)
2: Typical Urban Living	0.186103	9 (1.1)	284 (34.9)	96 (11.8)	75 (9.3)	265 (32.5)	12 (1.5)	74 (9.1)
3: Deprived	0.176744	2 (0.2)	11 (1.5)	0 (0)	8 (1.0)	227 (30.4)	416 (55.7)	82 (11.1)
4: Suburbia	0.277261	277	52 (7.7)	232	83 (12.1)	7 (1.1)	0 (0)	31 (4.6)
5: Outer Urban	0.175482	8 (9.8)	43 (51.1)	32 (38.9)	0 (0)	0 (0)	0 (0)	0 (0)
6: Young Urban Families	0.170864	2 (2.2)	88 (82.2)	3 (2.9)	1 (0.6)	8 (7.1)	0 (0)	5
Overall Performance		**129**	**100**	**105**	**89**	**138**	**143**	**90**

Table 3 Cross tabulation results

Table 3A: is Cross tabulations of the Clusters of ONS LSOA-level geodemographic classification Super-Groups and those of the alternative 'Geo-Social Area Classification' (GSAC) of England; 3B shows the index scores derived from the cross tabulation results. An index score of 100 means that the a cluster cell has the same population share of LSOAs, 0 means cluster is absent and 50 signifies a cluster is half as present and 200, twice as present , in that order

3.3 Validating the Classifications with Independent Datasets

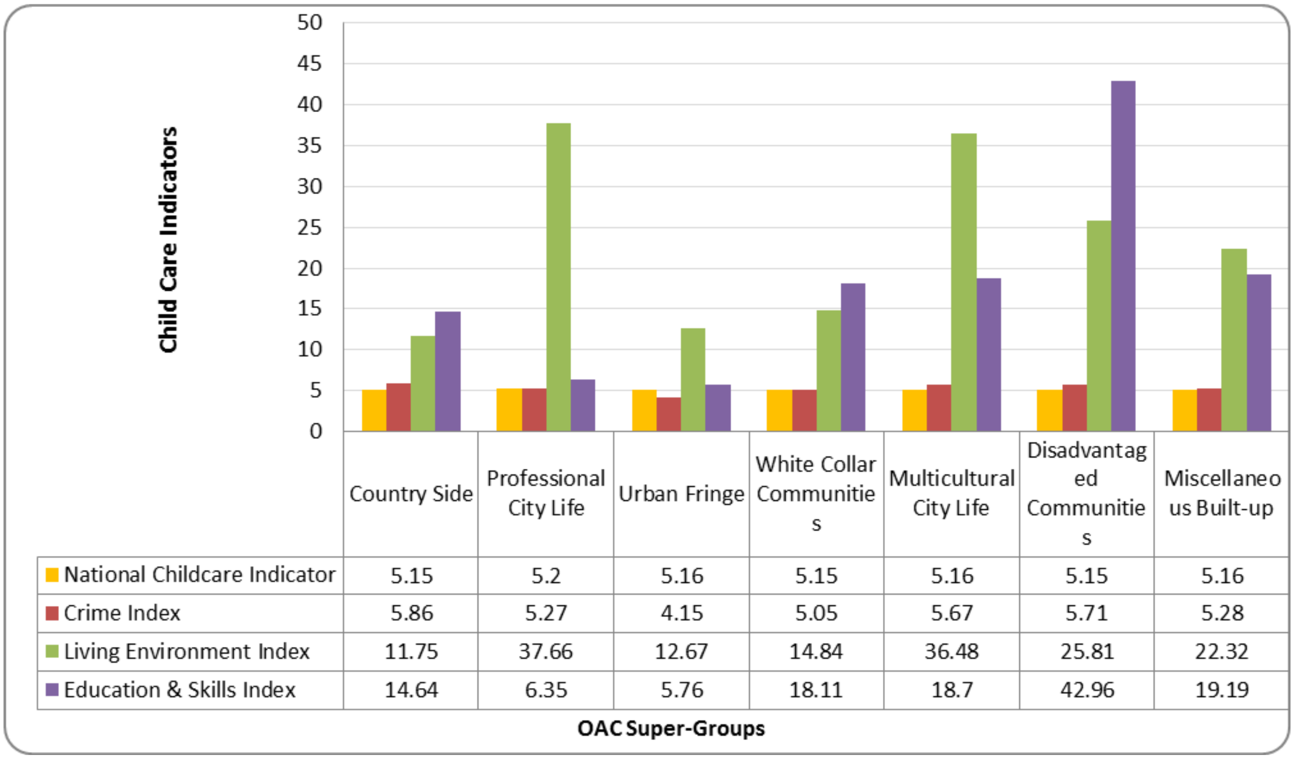
The GSAC and the ONS Super-Groups classification performance were examined in relation to the averages of the National Childcare Indicator data (NCI), three indicators used for the construction of the 2010 IMD namely Crime, Living Environment and Education and Training for each area type in London GOR. The results in Figures 2 and 3 show a clear stratification of these indicators along socioeconomic lines. In both classifications, higher proportions of low income groups in urban centres are more likely to take up the formal childcare element of the working tax credit (Gregory, 2009) in comparison with families in more rural areas and elderly populations in London. As expected, crime incidence is demonstrated to be relatively higher in socially disadvantaged and multicultural LSOAs of London located in communities such as Tower Hamlets and Westminster compared with suburban Greenwich and Barnet neighbourhoods. Poorer environmental conditions, educational qualification and professional skills are found to increase with area-level deprivation. Families in suburban communities are shown to live in better socioeconomic conditions. Though similar patterns of socioeconomic stratification are identified by both classifications, the ONS Super-Group which is constructed from a wider range of variables appears to illustrate small area-level socioeconomic stratification of areas more distinctively.



	Struggling Families	Typical Urban Living	Deprived Communities	Suburbia	Outer Urban	Young Urban Families
🟨 National Childcare Indicator	5.16	5.17	5.16	5.16	5.22	5.2
🟥 Crime Index	5.67	5.5	5.8	5.44	4.98	5.17
🟩 Living Environment Index	30.58	36.15	40.95	18.4	32.7	41.08
🟪 Education & Skills Index	30.59	13.99	36.57	7.34	3.3	12.29

Figure 2: Geo-Social Area Classification (GSAC) Clusters cross tabulated with other socioeconomic data not included in the classification

Lower values of National Child care indicator implies low take up of formal childcare tax credit and higher values refer to higher take up. Lower values of crime, living environment and education mean better conditions and higher values are indicative of poorer conditions



	Country Side	Professional City Life	Urban Fringe	White Collar Communitie s	Multicultural City Life	Disadvantag ed Communitie s	Miscellaneo us Built-up
🟨 National Childcare Indicator	5.15	5.2	5.16	5.15	5.16	5.15	5.16
🟥 Crime Index	5.86	5.27	4.15	5.05	5.67	5.71	5.28
🟩 Living Environment Index	11.75	37.66	12.67	14.84	36.48	25.81	22.32
🟪 Education & Skills Index	14.64	6.35	5.76	18.11	18.7	42.96	19.19

Figure 3: AC-Super Groups cross tabulated with other socioeconomic data not included in the classification

Lower values of National Child care indicator implies low take up of formal childcare tax credit and higher values refer to higher take up. Lower values of crime, living environment and education means better conditions and higher values are indicative of poorer conditions

3.4 Geodemographic Classifications and Health Inequality

The created classifications was also related to quintiles of health indicators from the 2001 census to examine how well health inequalities in England could be identified. The results show health outcomes to reflect the social characteristics of neighbourhoods and the people who live within them at the LSOA level. Table 4 shows the area correspondence of ONS Super-Groups and GSAC Clusters with LLTI and IB health measures for the year 2001. The cells with values highlighted in blue represent LSOA types with high proportions of good health outcomes. The cells in red are area types demonstrating high levels of ill-health. Areas highlighted in yellow represent the proportions of typical urban areas. The distribution of the quintile of health ratios across geodemographic clusters is better visualized in Figure 4, which clearly shows the more deprived/disadvantaged/multicultural areas as having poorer health outcomes (these are represented with red bars) compared with the affluent suburban groups with lower proportions of ill-health, depicted with blue bars. The presence of rural-urban differentials in health patterns is also clearly seen Core urban areas (ONS White Collar Communities and GSAC Typical Urban

Living groups) appear to reflect higher inequality. Most LSOAs in urban areas contain a complex mix of the first four illness quintiles almost in equal proportions compared to more rural ones. Note that the proportions of LSOAs within core urban neighbourhoods in the highest illness quantile (Q5) are relatively small compared with other neighbourhoods. These areas have high concentrations of younger professional adults who are less likely develop critical health issues. Overall, the results suggest that geodemographic classifications can be a more practical tool for explaining geographical variations in health.

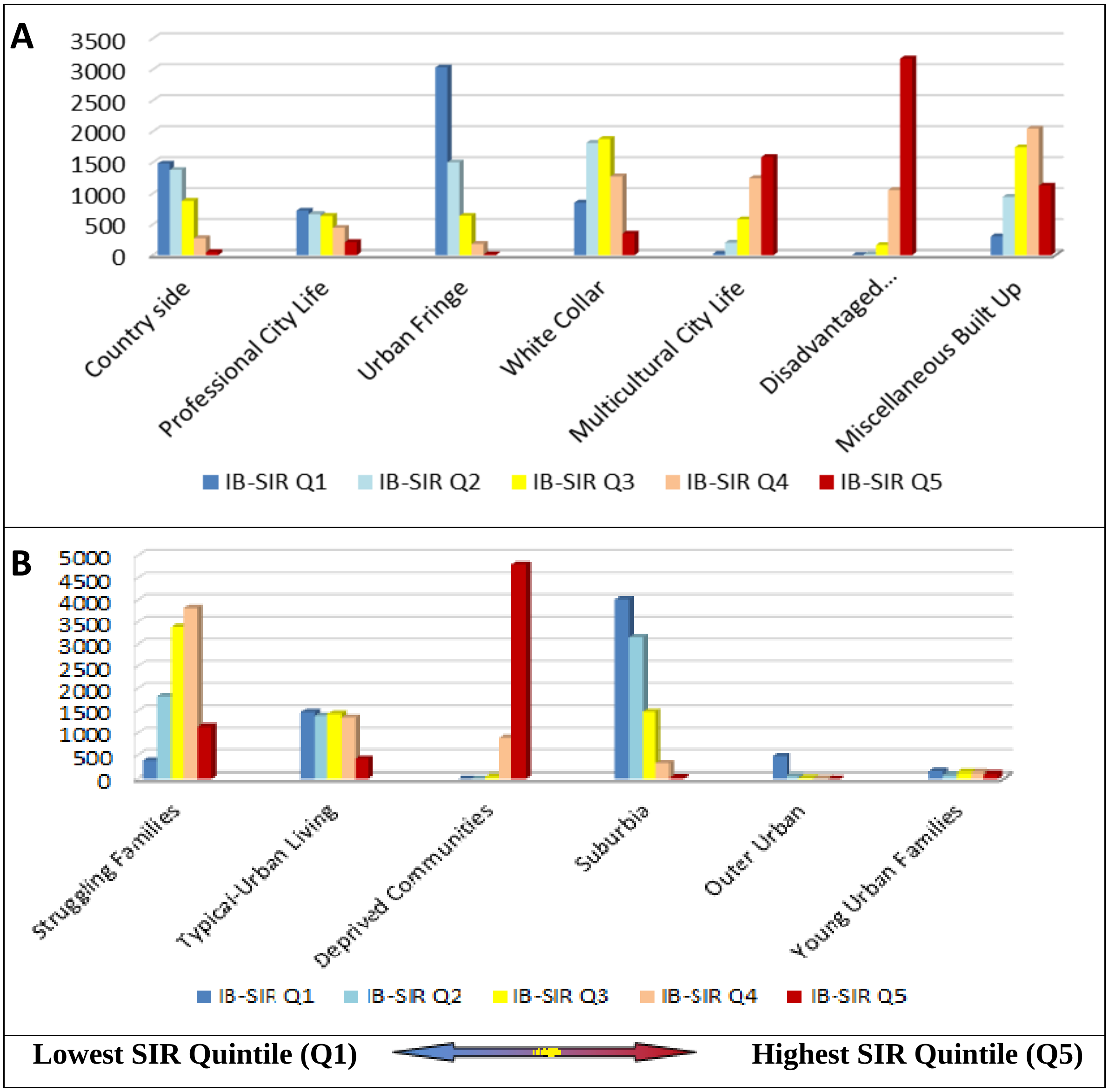


Figure 4: Quintiles of health measures by geodemographic typologies in England

The height of the bars represents the count of area types represented in a particular quintile of health. Blue bars represent better health (Q1 and Q2), red bars depict worse health (Q4 and Q5) and yellow bars are average health (Q2)

4.0 Conclusion

The challenges of data and methodological limitations of the K-Means clustering algorithm, the findings of the study demonstrate national administrative statistics used for creating the geodemographic classification to be of high performance given the strong associations between the datasets and equivalent census measures. The alternative area classification labelled ‘GSAC’ was found to stratify LSOAs into similar area types with the ONS Super-Groups. A high level of ecological correspondence was observed between the urban clusters of both classifications. Deprived communities of the GSAC area types appear to be clearly mapped out in a similar fashion with census definitions. This is expected given the heavy reliance of the classification on benefit data and council tax bands. The test of the classification against independent child health indicators for London not used in the both classifications further confirms the similarity of the GSAC with the OAC Super groups. In both classifications, poorer health, worse living environment index and higher crime rates are observed in more disadvantaged LSOAs while more affluent neighbourhoods record better health, improved living environment conditions and much lower records of crime.

References

Birkin, M. & Clarke, G. 1998. Gis, Geodemographics, And Spatial Modeling In The Uk Financial Service Industry. *Journal Of Housing Research,* 9**,** 87-111.

Birkin, M. & Clarke, G. 2009. Geodemographics. *The International Encyclopaedia Of Human Geography***,** 382-89.

Boyle, P., Norman, P. & Rees, P. 2004. Changing Places. Do Changes In The Relative Deprivation Of Areas Influence Limiting Long-Term Illness And Mortality Among Non-Migrant People Living In Non-Deprived Households? *Social Science & Medicine,* 58**,** 2459-2471.

Dedman, D., Hennell, T., Hooper, J. & Tocque, K. 2006. Using Geodemographics To Illustrate Health Inequalities. *Liverpool: North West Public Health Observatory, Liverpool John Moores University*.

Gregory, I. 2009. Childcare Takeup And National Indicator 118: A Summary Of Learning Funded By Government Regional Offices 08/09. *In:* Department For Children, S. A. F. (Ed.). London: Government Office For London. . [Accessed 20th August, 2012]. Available From: Http://Www.Daycaretrust.Org.Uk/Data/Files/Consultancy/Childcare_Take_Up_And_National_Indicator_118.Pdf.

Neighbourhood-Statistics 2004. Super Output Areas (Soas): Frequently Asked Questions. Office For National Statistics. [Accessed 20th July, 2012]. Available From: Http://Www.Neighbourhood.Statistics.Gov.Uk/Dissemination/Info.Do;Jessionid=Gqcqqlrlnvtgz1tgbflvnlthkvhy0cclhksyjgqj8fd1pv1d0gkd!1949496690!1342738827103?M=0&S=1342738827103&Enc=1&Page=Aboutneighbourhood/Geography/Superoutputareas/Soafaq/Soa-Faq.Htm&Nsjs=True&Nsck=True&Nssvg=False&Nswid=1366.

Ons 2011. Beyond The 2011 Census Project. Office For National Statistics. [Accessed 13th June, 2012]. Available From: Http://Www.Ons.Gov.Uk/Ons/About-Ons/What-We-Do/Programmes---Projects/Beyond-2011/Index.Html.

Petersen, J. 2009. *Social Marketing And Public Health.* Ucl (University College London).

Rogerson, P. A. 2010. *Statistical Methods For Geography: A Student's Guide*, Sage Publications Ltd.

Shelton, N. J., Birkin, M. H. & Dorling, D. 2006. Where Not To Live: A Geo-Demographic Classification Of Mortality For England And Wales, 1981-2000. *Health & Place,* 12**,** 557-569.

Vickers, D. & Rees, P. 2006. Introducing The Area Classification Of Output Areas. *Population Trend-London,* 125**,** 15.

Vickers, D. & Rees, P. 2007. Creating The Uk National Statistics 2001 Output Area Classification. *Journal Of The Royal Statistical Society: Series A (Statistics In Society),* 170**,** 379-403.

Voas, D. & Williamson, P. 2001. The Diversity Of Diversity: A Critique Of Geodemographic Classification. *Area,* 33**,** 63-76.

Webber, R. & Butler, T. 2007. Classifying Pupils By Where They Live: How Well Does This Predict Variations In Their Gcse Results? *Urban Studies,* 44**,** 1229-1253.