An Area Classification of the Jewish Population of England & Wales

Dissertation submitted for the degree of

MSc Social Research Methods

London School of Economics & Political Science

Daniel Paul Vulkan

Supervisor: Professor Jouni Kuha

August 2014

An Area Classification of the Jewish Population of England & Wales

ABSTRACT

National area classifications are now carried out as a matter of course after censuses in England and Wales. This project explores the feasibility of carrying out a similar exercise, focusing on a subset of that data – specifically, in this case, those people who chose to identify as Jewish by religion. If successful, the results may be of assistance to communal leaders in planning for the future.

*875 "Jewish Output Areas" (JOAs) were created, with roughly equal-sized Jewish populations. A set of 58 variables was selected, and a series of k-means cluster analyses were carried out. The adopted solution identified eight distinct clusters – Strictly Orthodox (*Haredi*); *Haredi*-Lite; Students; Older Communities; Young Cosmopolitan Professionals; Engaged Families; Disconnected; and On the Margins. The names reflect the dominant traits identified within the clusters.*

Improvements to the methodology have been suggested in order to produce JOAs which exhibit greater internal homogeneity.

INTRODUCTION

A population census of England & Wales has been carried out every ten years since 1801 (with the exception of 1941), with parallel censuses also being carried out in the other constituent countries of the United Kingdom. The aim of the census is to enumerate every individual and every household in the country, and to collect additional sociodemographic data which will help to inform public policy.

Topics covered have varied from census to census. In addition to basic demographic information, for example, the 2011 census questionnaire included questions on ethnicity, national identity, languages spoken, employment, education, health and accommodation, amongst others.

NATIONAL CLASSIFICATION OF CENSUS OUTPUT AREAS

Following every census since 1971, the Office for National Statistics (ONS) (or its predecessor the Office of Population Censuses and Surveys), the body responsible for carrying out the census, has produced a set of area classifications. These can be seen as a form of data reduction – an attempt is made to summarise all (or a large part) of the census data by assigning each geographical area to a single class, the members of which share similar characteristics.

The idea is not new. Between 1886 and 1903, the social reformer, Charles Booth, carried out an “Inquiry into the Life and Labour of the People of London”. One of the outcomes of his investigations was a series of maps of London districts, on which each street was colour-coded to represent a particular classification. His seven classes included categories with descriptions such as “Lowest class. Vicious, semi-criminal” and “Very poor, casual. Chronic want”, through to “Upper-middle and Upper classes. Wealthy”[1].

Following the publication of the results of the 2001 Census, the ONS carried out a number of exercises in order to produce area classifications for different geographies (http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/national-statistics-area-classifications/national-statistics-2001-area-classifications/index.html). In particular, Vickers et al (2005) were asked to produce a National Area Classification of the “Output Areas” (OAs), the lowest level of census geography, covering the whole of the United Kingdom.

After initially intending to use a combination of k-means clustering followed by hierarchical clustering (Ward’s method), it was concluded that this approach would not work, largely due to the presence of extreme outliers. Ward’s method on its own had been rejected due to the large number of data points present. The method finally adopted for this was k-means cluster analysis on its own, but carried out iteratively in order to produce a three-tier hierarchy, comparable to those produced for the other area classifications. The result was 7 clusters at the top layer, 21 at the second layer, and 52 at the lowest layer.

A similar methodology was adopted following the 2011 Census, although the technical notes (see ONS, 2014) are very sparse. The outcome this time was 8 “supergroups”, 26 “groups” and 76 “subgroups”.

Since the methodology adopted for this project draws very heavily on the 2001 and 2011 exercises, detailed discussion will be deferred to the Methodology section below.

[1] See the Charles Booth Online Archive, hosted by LSE (http://booth.lse.ac.uk/static/a/4.html).

THE RELIGION QUESTION

Apart from a "Religious Worship Census" that took place in parallel with the Census of 1851, the 2001 Census of England & Wales was the first to include a question on religion. The wording was "What is your religion?", with the following 8 options being offered, of which only one could be selected:

- None
- Christian (including Church of England, Catholic, Protestant and all other Christian denominations)
- Buddhist
- Hindu
- Jewish
- Muslim
- Sikh
- Any other religion

(the last of these requiring the religion to be written in).

Unlike every other question on the census form, this question was voluntary, and no values were imputed where this question was omitted. Most census output incorporating religion as a variable therefore used a 9-category breakdown, comprising the 8 categories above together with "Religion not stated".

The religion question was repeated for the 2011 Census, with almost identical wording.

THE BRITISH JEWISH COMMUNITY

The 2001 Census recorded 259,927 people in England and Wales who identified as Jewish, comprising around 0.5% of the total population. In 2011, this increased very slightly to 263,346, still 0.5%. Jews were recorded in each of the 348 local authorities in England and Wales, although very heavily concentrated in certain areas, including Greater London and its immediate surroundings (around two-thirds of the total), Greater Manchester and Leeds (see Graham et al, 2012).

The British Jewish community is far from monolithic. In particular, a full range of religious observance can be seen, from strict adherence to all aspects of Jewish religious law at one end of the spectrum, to people who are completely non-observant (but who may, for example, self-identify as "culturally Jewish"). Those who are more observant will have particular requirements which may influence their decisions on where to live. For example, they would normally wish to live within walking distance of a synagogue, since all forms of transport other than walking are prohibited on Saturday, the Jewish Sabbath. They might also, for example, want to live in an area where there are a sufficient number of observant Jews to justify facilities for the provision of kosher meat.

Graham et al (2007) noted that, in the 2001 Census, there were 116,330 households in England & Wales with a Jewish "Household Reference Person" (HRP)[2]. However, in only 89,371 of these (almost half of which were single-person households) were all members of the household identified as Jewish. There were a further 56,089 households in which at least one person, but not all, identified as Jewish, leading to a total of 145,460 households with at least one Jewish member. There is thus no straightforward definition of a "Jewish household", although Graham et al do suggest the figures of 116,330 and 145,460 could be referred to as "narrow" and "broad" definitions respectively. Figures collated by the Board of Deputies of British Jews (Graham et al, 2010) show that 86,115 households (across the whole of the UK, although numbers in Scotland and Northern Ireland are very low) were affiliated to or associated with synagogues in 2001. This gives an estimated "synagogue membership rate" (by household) of between 59% and 74%, depending on the definition of Jewish household. It is therefore clear that a very significant proportion of those identifying as Jewish were not associated with a synagogue.

(By 2010, the number of households affiliated to or associated with synagogues had fallen to 82,963. The number of households with a Jewish HRP recorded in the 2011 Census was 110,725; data on mixed households have not yet been published.)

A particular phenomenon of the Jewish community in the UK, and in other countries with significant Jewish populations, is the existence of a subgroup of "strictly Orthodox" Jews (also referred to in Hebrew as *haredi* – literally meaning "one who trembles [at the word of God]"). Those not falling in this group are sometimes described as "mainstream". There is however no agreed definition of these categories. The former can perhaps be characterised by strict adherence to Jewish religious law, combined with a rejection of what are considered to be morally harmful outside influences. Other traits often associated with the strictly Orthodox community are large average family sizes, and poor levels of secular education.

[2] Broadly speaking, the HRP in a multi-person household is the person with the highest economic status (in the priority order of full-time employed, part-time employed, unemployed, retired, other), with age being used as a tiebreaker.

Strictly Orthodox Jews tend to live in geographically highly concentrated areas, although in some places they live in areas which also have large "mainstream" Jewish populations and, in these cases, it is less easy to determine the geographic boundaries, if indeed they exist. The largest concentration of strictly Orthodox Jews is in Stamford Hill in North London, covering a large part of the northern end of the London Borough of Hackney, together with the south-east corner of Haringey. Other areas where strictly Orthodox Jews are assumed to make up the overwhelming majority of the Jewish population are Broughton Park in North Manchester (actually located just outside the City of Manchester, in the north-eastern corner of the neighbouring City of Salford), and the Bensham district of Gateshead. Golders Green, Hendon and Edgware (all in the London Borough of Barnet), and Sedgley Park (at the southern end of the Metropolitan Borough of Bury) are all thought to have mixed populations.

Research into the strictly Orthodox community has been carried out, amongst others, by Holman et al (2002) and Vulkan et al (2008). It is estimated that the strictly Orthodox currently comprise between 12% and 15% of the total British Jewish community, and are growing in absolute terms at a rate of around 4% per annum.

It is, of course, not known how many Jews chose not to answer the religion question on the census form. A number of Jewish community surveys conducted after the 2001 Census suggested that the overall non-response rate among Jews to the question was either very similar to that of the general population (around 7.7%), or slightly higher. It is conceivable that memories of persecution (particularly in continental Europe during the Second World War) would make some Jews more reluctant than the general population to declare their religion "openly" (although the census data are of course held securely).

Additionally, some Jews hold that there is a religious prohibition on "counting" Jews. It is understood that, at the time of the 2001 Census, many rabbinic leaders in the strictly Orthodox community advised their followers not to answer the religion question for this reason. However, some of them appear to have reversed their advice for the 2011 Census.

RESEARCH QUESTION

The purpose of this project is to carry out an analysis, along the lines of that carried out by Vickers et al and by ONS, on the subset of the 2011 Census data consisting of those individuals who ticked the "Jewish" box in response to the religion question. The intention is to establish whether such an analysis is feasible and, if so, if the results produced are credible (ie whether or not they are consistent with other information that is already known about the Jewish communities in particular areas).

The results of the analysis may prove useful to community leaders, for example by identifying areas which may have similar communal needs.

The possibility of identifying the strictly Orthodox population is of particular interest. It is expected that some indicators will emerge from the cluster analysis, and may help to identify localised *haredi* pockets within mixed communities.

There may be opportunities for subsequent qualitative research (eg interviewing individual Jews and/or Jewish community leaders) in order to investigate the reasons for any patterns that are identified as a result of the quantitative research – particularly if these are not in line with expectations.

Whilst this paper focuses on the Jewish community, with which the author is most familiar, similar techniques should be applicable to other religious minority communities.

METHODOLOGY

GEOGRAPHY

In 1991, the smallest geographical area for which census results were published was the Enumeration District (ED), which also served as the base unit for data collection (which was carried out door to door). EDs had been constructed manually, and were constrained to fit exactly within administrative boundaries (wards and local authorities). On analysis, however, it was found that EDs varied considerably in terms of the number of households and persons which they contained. In many cases, these fell below the thresholds for publication, and EDs had to be combined. The EDs also lacked social homogeneity.

A new approach was adopted for the 2001 Census, drawing on the automated zoning procedure (AZP), originally devised by Openshaw (1977), and subsequently developed by Openshaw and others. Its application to the 2001 Census was outlined by Martin et al (2001) and Harfoot et al (2010). The smallest reporting unit was the Output Area (OA). Broadly speaking, these were constructed by aggregating groups of one or more adjacent "building blocks", comprising unit postcodes.

After an initial random aggregation (constrained only by minimum thresholds for households and head counts, and the requirement for OAs to be completely nested within administrative wards), the quality of each OA was assessed by examining (i) how far the population size of that OA diverged from the target population size; (ii) the geographical compactness; and (iii) the degree of social homogeneity, as measured in this instance by reference to tenure and accommodation type. Following this, all possible swaps of building blocks between OAs were considered in turn, to determine whether this improved the overall quality. The process was then repeated from the initial random aggregation stage, for a total of 100 iterations, and the iteration producing the best quality was chosen.

The result was 175,434 OAs in England and Wales, with a mean population of 297 and mean number of households of 124. The minimum threshold was 100 people or 40 households. Larger reporting areas were also created by aggregating OAs into 34,378 LSOAs (lower layer super output areas) and 7,194 MSOAs (middle layer super output areas).

For the 2011 Census, the Office for National Statistics (ONS) decided to retain the 2011 OAs as far as possible in order to aid comparability. OAs were however split or merged where (i) there had been significant population change; (ii) boundary changes resulted in OAs straddling a local authority boundary; or (iii) it was determined that the 2001 OAs had in fact lacked social homogeneity. This process was carried out by the Census2011Geog project at the University of Southampton (http://census2011geog.census.ac.uk/).

As a result, the number of OAs in England and Wales increased to 181,408, with a mean population of 309 (median 303), and a mean number of households of 219. Overall, around 2.6% of the 2001 OAs were affected by changes (Tait, 2012). The total number of people recorded in each OA varied between 91 and 4,140, and 90% of all OAs had populations in the range 198 to 438.

Both Vickers et al (2005) for the 2001 Census, and ONS (2014) for the 2011 Census, used OAs as the units for their area classifications.

It was intended that a similar approach be used for this project. However, the number of Jews recorded in each OA in 2011 varied between 0 and 445, with a mean of 1.45. No Jews at all were recorded in 132,390 of the 181,408 OAs (ie 73%), hence the median was 0. Excluding those OAs with no Jews, the mean was 5.4 and the median was 2. In 95% of OAs, there were 17 Jews or fewer, and a majority of all Jews recorded were located in a total of just 1,639 OAs (0.1% of the total number of OAs).

This is a highly skewed distribution, comprising a very large number of tiny values, and a small number of very large values. Any form of cluster analysis which relied on OAs as the data points was therefore going to run into severe problems at an early stage.

The solution was to aggregate OAs into larger units, before any analysis is attempted. The same method that was used to derive the OAs themselves in 2001 (the automated zoning procedure as described above) would be used to create a number of "Jewish Output Areas" (JOAs), which would be contiguous, and of comparable size in terms of their Jewish population. The target was for a mean Jewish population of around 300 (ie similar to the mean total population of the 2011 Census OAs) which, given a total Jewish population of 263,346, would suggest around 880 JOAs.

Initial attempts at this, using the AZTool software (see References), and using OAs as the building blocks, failed to produce any solutions. The approach adopted by the software, of allocating the building blocks to "tracts" at random and then iteratively swapping blocks between tracts to improve the solution in line with the targets set, proved unable to deal with the large number of blocks with a value of zero (ie OAs with no Jews).

The use of the larger MSOAs was then considered as an alternative. Here, the number of Jews recorded in the 7,201 MSOAs in England and Wales varied between 0 and 5,574, with a mean of 37 and a median of 7. Although 471 MSOAs (6.5% of the total) recorded no Jews, it was hoped that the software would be better able to deal with this. However, a problem now occurred at the other end of the scale, with 147 MSOAs exceeding the target size of 300. These MSOAs contained 151,592 Jews (57.6% of the total). Being forced to use these undivided would have prevented any possibility of examining what might be happening at a very local level.

The solution adopted was to use MSOAs where the number of Jews was below a certain threshold, but to split larger MSOAs into their component OAs. The threshold was set somewhat arbitrarily at 500 (meaning that a resulting JOA could still be as large as 499), and a total of 99 MSOAs were found to exceed the threshold. By coincidence this produced a fairly even split, with the 99 "large" MSOAs containing a total of 133,647 Jews, leaving 129,699 in the remaining 7,102 "small" MSOAs. The large MSOAs were split into a total of 2,497 OAs, giving an overall total of 9,599 building blocks.

For simplicity, the MSOA building blocks and OA building blocks were processed separately. This had the effect of introducing a constraint, that any individual JOA could not comprise a combination of OAs and MSOAs.

Boundary files (in the standard ESRI shapefile format) were obtained from the ONS for all MSOAs (see References), and pre-processed using the AZTImporter software to create the input text files necessary for AZTool (one containing attribute data for each MSOA, and the other a list of boundaries between MSOAs). The large MSOAs were manually removed.

A problem encountered when first processing the MSOAs was the presence of two "islands" (literally). The Isles of Scilly (45km off the coast of Cornwall, with a total Jewish population of 4) and Anglesey (adjacent to the North Wales coast, with a Jewish population of 40) were each much too small to comprise JOAs in their own right, but were not physically adjacent to any other MSOAs, due to the intervening sea. For the purpose of creating the JOAs, these populations were added to MSOAs located on the mainland.

Parameters are supplied to AZTool in the form of an XML text file – the final version used for processing the MSOAs appears as Item 1 in the Appendix. A target size of 300 was set, with a minimum threshold of 100 and a maximum threshold of 500. Tracts were constrained not to cross boundaries between regions (ie the nine former "Government Office Regions" in England, plus Wales). The program was set to run 20 times, with 20 swap iterations in each run, and the "best" run was saved.

A parallel process was then carried out with the OAs. A total of 875 JOAs were produced, with a mean size of 301.0 and median 294. The smallest JOA (consisting of the Isle of Wight) contained 124 Jews, with the next smallest being 160; the largest was 494.

QGIS (see References) was used to combine the OA and MSOA boundaries into JOA boundaries, and names were allocated to each JOA, comprising "J" followed by two digits denoting the region and three digits allocated in sequence within each region. The result is shown in Figure 1 overleaf.

It will be seen that the JOAs vary tremendously in geographical size. The largest, J00002, has an area of 9,688 km^2, covering almost half of the total area of Wales. The smallest, J07265, is roughly a millionth of the size at 0.01 km^2, comprising less than a single block in the north of the London Borough of Hackney (see Figure 2).

It is worth noting, in passing, that other census geographies were also considered as an alternative to OAs and MSOAs. In particular, there are 5,493 "Built Up Areas" (BUAS) in England and Wales, with total populations varying between 101 and just under 10 million (Greater London). Jews are present in 2,961 of these, with numbers varying between 1 and just under 170,000. Whilst 95% of the total population of England and Wales live within Built Up Areas, almost 98.5% of Jews do. It would, however, have been necessary to exclude the smaller areas in order to carry out a cluster analysis; and it would not have been possible to investigate individual neighbourhoods within the major conurbations.

Figure 1: JOAs (Jewish Output Areas) in England & Wales

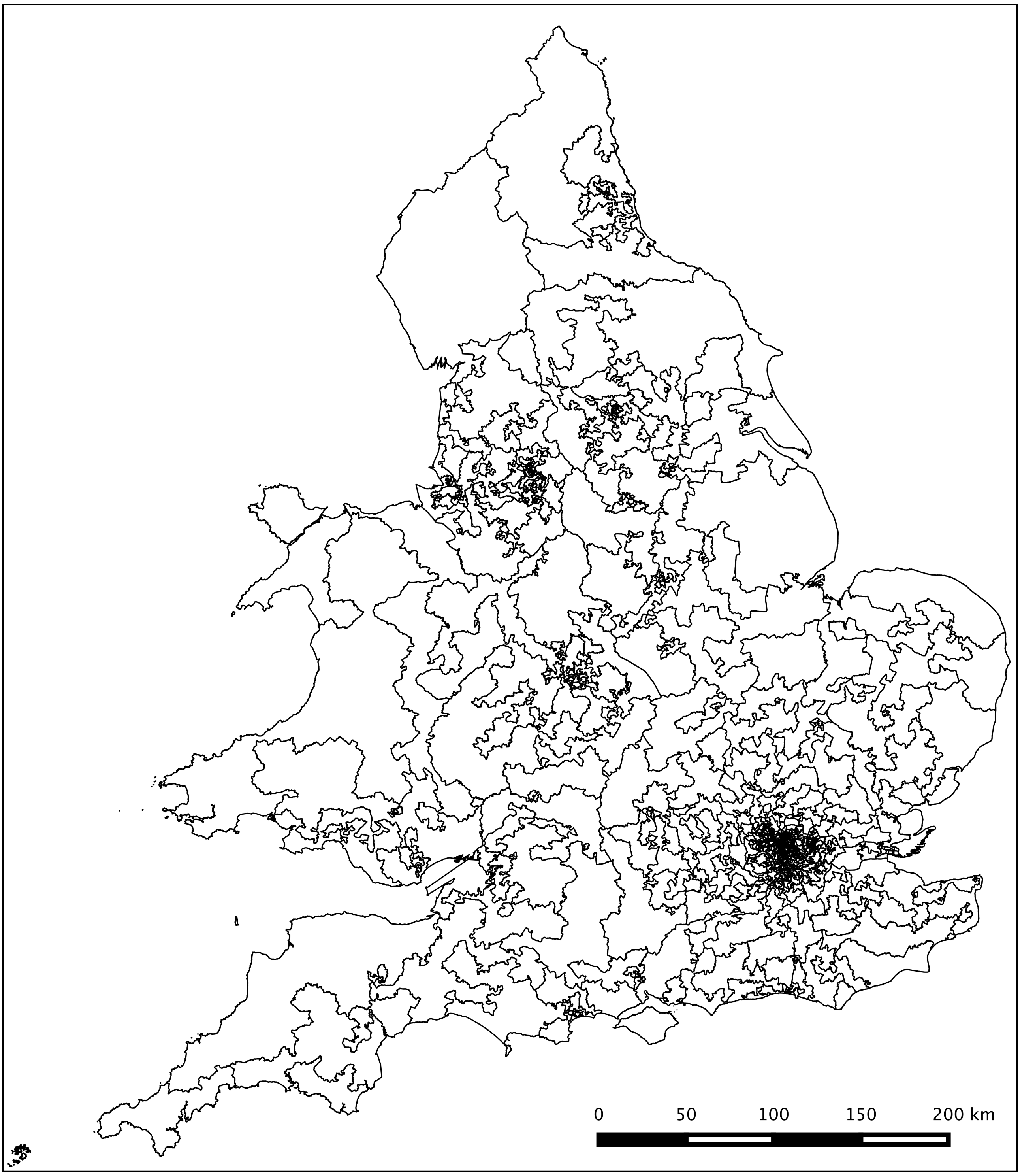
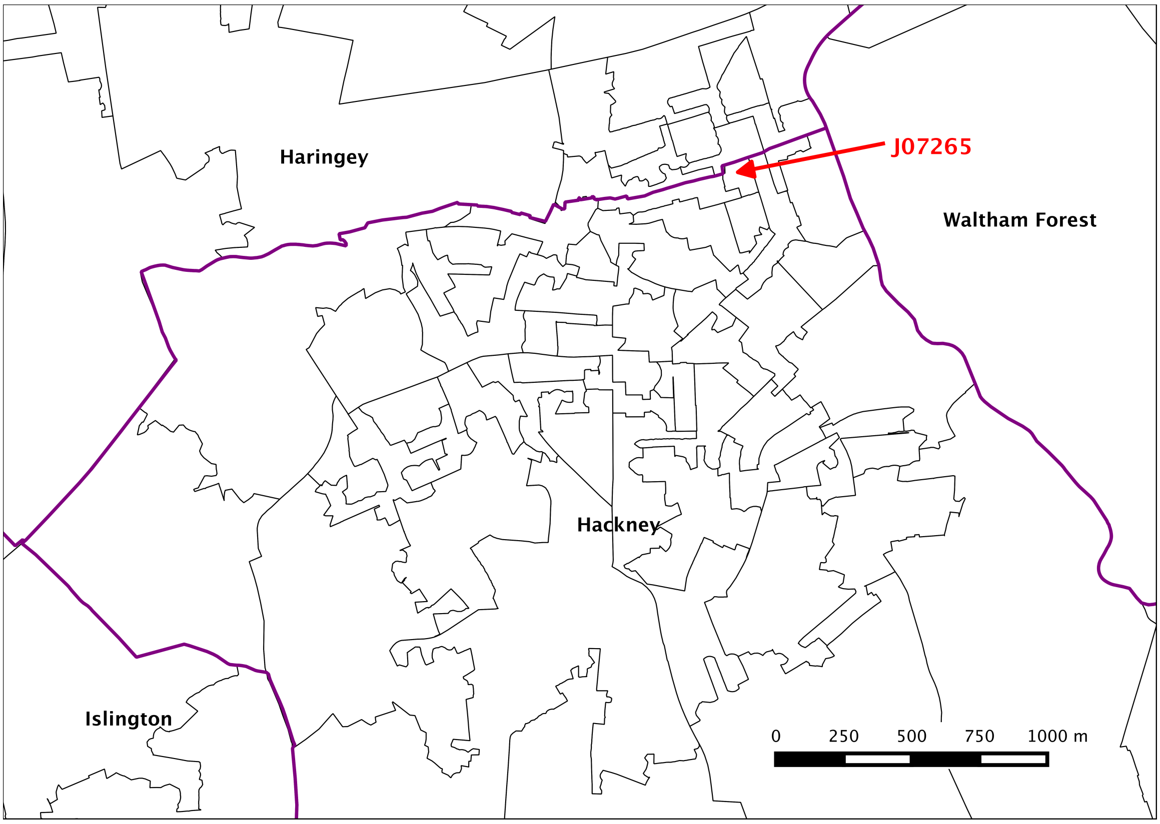


Figure 2: Close-up of Stamford Hill, showing smallest JOA (J07265)



VARIABLE SELECTION

There is no objective, mechanical method for determining which subset of available variables should be used in a cluster analysis. Rejecting a particular variable can be thought of as equivalent to attaching to it a weight of zero.

Vickers et al (2005) started with the data available from the "Key Statistics" tables, which were the first results to be available at the Output Area level. These were primarily univariate tables (ie tabulating the counts for the possible values of one variable only for each geographic area), although a small number also included a breakdown by sex. They quickly reduced the full list of around 350 variables to 94, by aggregating some values within tables; combining results for males and females; and discarding some variables which were not defined consistently across the whole of the UK.

This list was then reduced to a final list of 41 variables, with the intention of including variables which would cover five main "domains" that had been identified – demographic structure, household composition, housing, socio-economic and employment. The methods used to choose the final 41 are described in detail by Vickers et al (pp6-16). They included:

- elimination of one of each pair of very highly correlated variables, where that correlation was either due to a shared denominator (eg there will inevitably be high negative correlation between the percentage of a population who are married and the proportion who are single), or inherent causality;
- further aggregation of values within tables which behave similarly or represent very small proportions of the population (eg combination of separated and divorced people, and of different types of flats);
- dropping variables which showed very little variation.

ONS (2014) considered both the "Key Statistics" and "Quick Statistics" tables published for the 2011 census. Starting with a long list of 167 variables, they ended up with a final list of 60 variables. Whilst not providing any detail of how the final selection was carried out, the five "domains" identified for the 2001 census were retained, and they state that the remaining variables are those considered "most important" for the purpose of the area classifications (p4).

For the purpose of selecting variables for this project, account was taken of the work of Vickers et al and the ONS, both in terms of the methods, and the actual variables chosen. The comments below outline those areas where it was necessary to diverge significantly from these. The full list of variables selected is set out in Table 1 below. Unless otherwise stated, variables are proportions of the total Jewish population (ie those who answered "Jewish" in response to the religion question).

Table 1: Variables used in cluster analysis

Variable name	Description	Source table
Demographic		
V01	Religion: Jewish (population base is total population)	KS209EW
V02	Religion: did not answer (population base is total population)	KS209EW
V03	Total population / total land area in hectares	KS209EW/ shapefiles
V04	Aged 0 to 15	LC2107EW
V05	Aged 16 to 24	LC2107EW
V06	Aged 25 to 34	LC2107EW
V07	Aged 35 to 49	LC2107EW
V08	Aged 50 to 64	LC2107EW
V09	Aged 65 to 74	LC2107EW
V10	Aged 75 and over	LC2107EW

Variable name	Description	Source table
V11	Ethnic group: White ▸ English/Welsh/Scottish/Northern Irish/British	LC2201EW
V12	Ethnic group: White ▸ any other (write in)	LC2201EW
V13	Ethnic group: Other ethnic group ▸ Arab OR any other (write in)	LC2201EW
V14	National identity: British (only)	LC2204EW
V15	National identity: English OR Welsh (only)	LC2204EW
V16	National identity: Other (write in) (with or without one or more UK identities)	LC2204EW
V17	Country of birth: UK OR Ireland	LC2207EW
V18	Country of birth: other countries in Europe	LC2207EW
V19	Country of birth: countries in Africa	LC2207EW
V20	Country of birth: countries in the Middle East or Asia	LC2207EW
V21	Country of birth: countries in the Americas	LC2207EW
Household composition		
V22	Household type: one person household (population base is all Jewish HRPs)	LC1202EW
V23	Household type: one family only, all aged 65 or over (population base is all Jewish HRPs)	LC1202EW
V24	Household type: one family only, married, same-sex civil partnership or cohabiting (population base is all Jewish HRPs)	LC1202EW
V25	Household type: one family only, lone parent (population base is all Jewish HRPs)	LC1202EW
V26	Household type: other household types (population base is all Jewish HRPs)	LC1202EW
V27	Living in a communal establishment	LC2120EW
V28	Number of Jews / number of Jewish HRPs	KS209EW/ LC1202EW
Housing		
V29	Living in a detached house (population base is all usual residents in households)	LC2120EW
V30	Living in a semi-detached house (population base is all usual residents in households)	LC2120EW
V31	Living in a terraced house (population base is all usual residents in households)	LC2120EW
V32	Living in a flat, maisonette, apartment, or mobile or temporary structure (population base is all usual residents in households)	LC2120EW
V33	Tenure: owned outright (population base is all usual residents in households)	LC4417EW
V34	Tenure: owned with a mortgage (population base is all usual residents in households)	LC4417EW
V35	Tenure: social rented (population base is all usual residents in households)	LC4417EW
V36	Tenure: private rented or living rent free (population base is all usual residents in households)	LC4417EW
V37	Occupancy rating (rooms): -1 or less	LC4207EW
Socio-economic		
V38	Age-standardised ratio for limiting long-term illness	LC3207EW
V39	Car or van availability: none (population base is all usual residents in households)	LC4417EW
V40	Car or van availability: one (population base is all usual residents in households)	LC4417EW
V41	Car or van availability: two or more (population base is all usual residents in households)	LC4417EW
V42	Highest level of qualification: none (population base is all usual residents aged 16 and over)	LC5204EW
V43	Highest level of qualification: level 1 (population base is all usual residents aged 16 and over) (1-4 “O”-Levels / CSEs / GCSEs – any grades – or equivalent)	LC5204EW
V44	Highest level of qualification: level 2 (population base is all usual residents aged 16 and over) (5+ “O”-Level Passes / CSEs Grade 1 / GCSE Grades A*-C or 1 “A”-Level or equivalent)	LC5204EW
V45	Highest level of qualification: level 3 (population base is all usual residents aged 16 and over) (2+ “A”-Levels or equivalent)	LC5204EW
V46	Highest level of qualification: level 4 (population base is all usual residents aged 16 and over) (Degree, professional qualification or equivalent)	LC5204EW
V47	Highest level of qualification: other (population base is all usual residents aged 16 and over) (Apprenticeship, vocational/work-related qualifications / overseas qualifications)	LC5204EW

Variable name	Description	Source table
Employment		
V48	Economic activity in previous week: working part-time (employee or self-employed) (population base is all usual residents aged 16 and over)	LC6205EW
V49	Economic activity in previous week: working full-time (employee or self-employed) (population base is all usual residents aged 16 and over)	LC6205EW
V50	Economic activity in previous week: student (employed, unemployed or economically inactive) (population base is all usual residents aged 16 and over)	LC6205EW
V51	Economic activity in previous week: unemployed (excluding full-time students) (population base is all usual residents aged 16 and over)	LC6205EW
V52	Economic activity in previous week: retired (population base is all usual residents aged 16 and over)	LC6205EW
V53	National Statistics Socio-economic Classification: 1.1 Large employers and higher managerial and administrative occupations 1.2 Higher professional occupations 2. Lower managerial, administrative and professional occupations (population base is all usual residents aged 16 and over)	LC6207EW
V54	National Statistics Socio-economic Classification: 3. Intermediate occupations 4. Small employers and own account workers 5. Lower supervisory and technical occupations (population base is all usual residents aged 16 and over)	LC6207EW
V55	National Statistics Socio-economic Classification: 6. Semi-routine occupations 7. Routine occupations (population base is all usual residents aged 16 and over)	LC6207EW
V56	Industry: G, I Distribution, hotels and restaurants (population base is all usual residents aged 16 and over in employment the previous week)	LC6212EW
V57	Industry: K, L, M, N Financial, Real Estate, Professional and Administrative activities (population base is all usual residents aged 16 and over in employment the previous week)	LC6212EW
V58	Industry: O, P, Q Public administration, education and health (population base is all usual residents aged 16 and over in employment the previous week)	LC6212EW

Since it was that subset of the population who chose to identify as Jewish that was to be considered, the univariate Key Statistics and Quick Statistics tables would have been of little use (except as noted below). The starting point instead was the "Local Characteristics" tables, each of which tabulate two or more variables, at geographies down to Output Area. Specifically, variables were selected from those tables which included religion as one of the tabulated variables. The tables were obtained through the Nomis website (see References).

(The "Detailed Characteristics" tables, which include finer categorisations of variables but only down to the MSOA geography, were considered, but rejected due to the complexity, and additional assumptions, that would have been required in order to estimate relevant values at smaller geographies.)

Vickers et al (2005), followed by ONS (2014), made no use of the religion question. Vickers et al implied (p12) that the main reason for this was that the question was voluntary in England and Wales, but compulsory in Scotland and Northern Ireland. (This was not in fact correct – the question was voluntary in all parts of the UK, in both 2001 and 2011, although it was only the England and Wales questionnaire that noted this explicitly alongside the question.)

Most of the variables considered in this project are, as would be expected, taken solely from the sub-tables comprising people who ticked "Jewish" in response to the religion question. However, it was considered that the level of "Jewishness" of an area (ie the relative size of the Jewish population to that of the population as a whole) might be a significant factor. The proportion of people choosing not to respond to the religion question was also retained since, as noted previously, it has been suggested that response rates may have been significantly lower, in particular, amongst strictly Orthodox Jews. These variables were derived from Key Statistics tables.

Dividing the total population in an area (excluding those living in communal establishments) by the number of "Household Reference Persons" (HRPs) should give a figure for average household size. However, this does not work for subsets of the population (such as those of a particular religion):

- not everyone living in a household where the HRP is Jewish will themselves necessarily be Jewish (or have chosen to identify as Jewish);
- not everyone who is Jewish will live in a household where the HRP is Jewish (or chose to identify as Jewish).

In fact, there is no such thing as the "average Jewish household size" since, as discussed previously, there is no obvious definition of "Jewish household". Nevertheless, it was felt that this ratio would still prove to be a useful proxy indicator. References later on to "household size" should be interpreted accordingly.

DATA PREPARATION, TRANSFORMATION AND STANDARDISATION

Vickers et al carried out three stages of data manipulation on the selected variables before carrying out the cluster analysis. ONS (p3) subsequently identified these stages as data preparation, data transformation and data standardisation.

Preparation

Vickers et al simply converted most variables into percentages, by dividing the counts by the relevant populations and multiplying by 100. The exceptions were population density (calculated as persons per hectare), and limiting long-term illness (LLTI), for which a standardised illness ratio (SIR) approach was adopted, as described on pp15-16. This was done in order to control for age – it would not make sense to compare unadjusted illness rates in areas with different age profiles, since older people are intrinsically much more likely to experience illness.

The SIR for an area *i* was calculated as:

$$SIR^i = 100 \times \frac{I^i}{\sum_a r_a^n P_a^i}$$

where I^i is the count of ill people in area *i*, r_a^n is the average rate of illness for age group *a* across the population as a whole, and P_a^i is the population in area *i* in age group *a*. The resulting figure will be a positive number, which may be greater or less than 100, with a value of 100 denoting the same level of illness that would have been expected if whole-population illness rates had been applied to the specific age structure of the area.

ONS considered the rates approach in addition to two others (neither of which they describe in detail) – calculating index scores, and calculating mean differences.

For this project, the approach adopted by Vickers et al was followed (this was also the approach ultimately adopted by ONS), including the use of SIR for LLTI.

Transformation

Outliers (ie extreme values) will result in greater weighting being implicitly applied to those variables, increasing the likelihood of some very small clusters (ie comprising just those data points exhibiting the extreme values). Skewness of the distribution of a variable can also have the same effect.

There are a number of functions which can be applied to the observed values of a variable in order to reduce this effect. ONS cite three:

- logarithmic transformation ($y=\log_e(x+1)$ – the "+1" being required to avoid the problem of the log function being undefined for 0);
- inverse hyperbolic sine transformation ($y=\sinh^{-1}(x)$);
- "Box-Cox" transformation.

Vickers had used the first of these, and ONS adopted inverse hyperbolic sine. For the purpose of this report, log and inverse hyperbolic sine were both considered, together with the option of leaving the data untransformed.

STANDARDISATION

Standardisation is the process of adjusting variables so that they are all expressed on the same scale. The two most common methods for this are:

- Z-score standardisation – measuring the distance of each observed value of a variable from the variable's mean, and then dividing by the standard deviation;
- range standardisation – a linear transformation, mapping the full observed range of the variable on to the range from 0 to 1.

A modification of the latter, described by ONS (although they ultimately opted for the unmodified version), is inter-decile range standardisation, in which the transformation is based on the range between the 10^{th} and 90^{th} percentiles of the observed values rather than the full range; this avoids the problem of values in the central part of the range being "squashed together" by the presence of outliers.

ONS justify the use of standardisation by stating (p3):

> "In general, variables with larger values and greater variation will have more impact on the final similarity measure. It is necessary to ensure each variable is equally represented in the distance measure by standardising the data."

However, it can be argued that this is not necessarily a desirable feature. As Everitt et al (2001) point out (p67):

> "when standardizing variables prior to analysis the investigator assumes that the importance of a variable decreases with increasing variability"

or, to put it another way, standardisation applies a greater weighting (compared to the unstandardised data) to those variables which show the least variation initially. In many cases, due to the lack of variation, these variables will not be particularly interesting. A slightly different method of variable selection may have resulted in these variables being omitted entirely (ie being given a weighting of zero). In a situation such as this, where (almost) all variables have already been transformed into rates (ie constrained to lie between 0 and 1), the case for standardisation seems weak.

Having said that, a more convincing case for standardisation can be made where the possible values of different variables are genuinely on different scales. This would apply, for example, to average household size, to any measures which are ratios (such as the SIR for LLTI) and, most significantly, to population density (which, in this dataset, ranges between 0.635 and 393.2).

It was therefore decided to consider three scenarios – no standardisation; range standardisation of all variables; and range standardisation only of those variables (V03, V28 and V38) which had not already been converted to rates.

CARRYING OUT THE CLUSTER ANALYSIS

Vickers et al chose, for the 2001 Census analysis, to apply a log transformation, followed by range standardisation, to all variables.

ONS, for the 2011 Census, produced 27 separate datasets, based on three options for each of the preparation, transformation and standardisation stages.

For this exercise, it was decided to produce nine datasets – three methods of transformation (none, logarithmic and inverse hyperbolic sine), combined with three approaches to standardisation (none, full, and partial, as explained above).

For each of these datasets, a series of nine k-means cluster analyses were then carried out using Stata, representing every possible number of clusters between one (trivial) and nine. Default options were retained for these calculations (the use of Euclidean distance to measure the similarity between data points, and a maximum of 10,000 iterations).

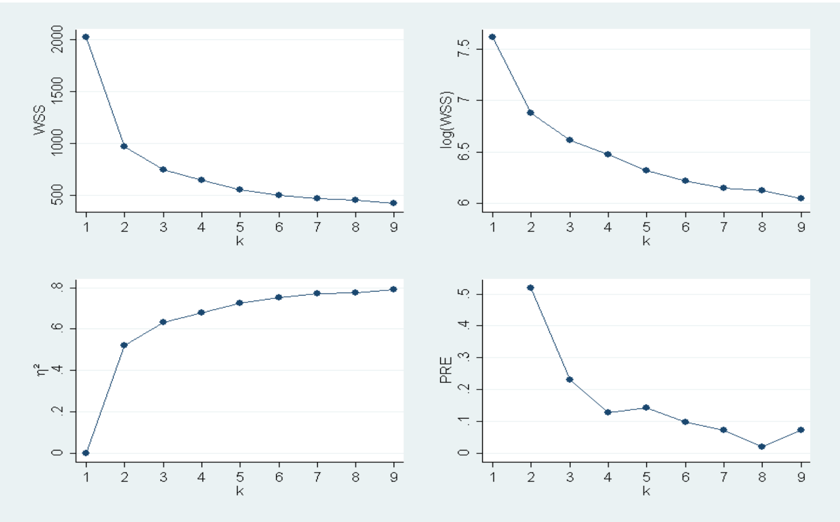
Stata does not readily produce any indicators which can be used to compare different outcomes of k-means cluster analysis. However, a method for doing so is suggested by Makles (2012). This involves calculating the "within sum of squares" (WSS) for each of the possible solutions (ie different numbers of clusters), and looking for a "kink" in the curve showing the reduction in WSS as the number of clusters increases. Log(WSS) is an alternative measure that may be used, as are η_k^2 (defined as 1 – WSS(k)/WSS(1)) and the proportional reduction of error coefficient, PRE_k = [WSS(k-1) – WSS(k)]/WSS(k-1).

Makles provides a Stata script to carry out these calculations and produce illustrative graphs. An adapted version of this script was used, and appears as Item 2 in the Appendix. Table 2 and Figure 3 below show the statistics generated for the dataset that was eventually chosen.

Table 2: Optimisation statistics – Inverse Hyperbolic Sine transformation, no standardisation

WSS[9,5]					
	k	WSS	log(WSS)	eta-squared	PRE
r1	1	2020.2003	7.610952	0	.
r2	2	969.44266	6.8767213	0.52012548	0.52012548
r3	3	745.47277	6.6140186	0.63099067	0.23102954
r4	4	649.70133	6.4765128	0.67839757	0.12847074
r5	5	556.21622	6.3211571	0.72467274	0.14388937
r6	6	502.18832	6.2189752	0.75141658	0.09713471
r7	7	466.00723	6.1442011	0.76932623	0.07204686
r8	8	456.86827	6.1243951	0.77385002	0.01961119
r9	9	423.23283	6.0479225	0.79049957	0.07362174

Figure 3: Optimisation statistics – Inverse Hyperbolic Sine transformation, no standardisation



The "optimal" solution was in fact chosen somewhat subjectively. Each of the 72 solutions generated (three methods of transformation, three methods of standardisation, and between 2 and 9 clusters) was viewed in

QGIS, firstly to see if it was "correctly" able to identify the distinctive nature of Stamford Hill (ie a clear distinction between those JOAs located in those areas "known" to be *haredi*, and those in the surrounding non-*haredi* areas). This immediately reduced the number of potential solutions to 42.

Other geographical areas were also examined, to look for anything obviously "wrong" (based on the author's knowledge of the areas concerned). Those solutions using an inverse hyperbolic sine transformation, and no standardisation, seemed generally suitable. The absence of even partial standardisation (ie of those variables which are not rates) was justified on the basis that the inverse hyperbolic sine transformation had already reduced the range of those variables considerably. For example, following the transformation, the maximum value of population density was reduced from 393.2 to 6.67.

From within this group, the solution with eight clusters was selected. This would not be considered the optimal solution based on Table 2 and Figure 3 above; the improvement between eight clusters and nine clusters looks to be greater than that between seven and eight. However, there is not much difference between any of these solutions, and the eight-cluster solution exhibited an interesting phenomenon of two distinct cluster types present within Stamford Hill (whilst also remaining distinct from the surrounding areas), which was felt to be worth investigating.

In considering the results presented below, it should therefore be borne in mind that some of the patterns present are the consequence of subjective decisions that were made by the investigator, rather than having emerged entirely naturally.

RESULTS

CLUSTER PROFILES

The distinctive characteristics of the eight clusters identified are discussed below. Table 3 presents some summary statistics; a full table of the values of the variables is included as Item 3 in the Appendix.

The names given to the clusters are intended to be simple reminders of the dominant features; it should be noted that the populations within the clusters are not of course homogeneous. It should also be remembered that all comparisons are within the subset of people who chose to identify as Jewish in the census, and not with the wider population.

The description of each cluster is preceded by:

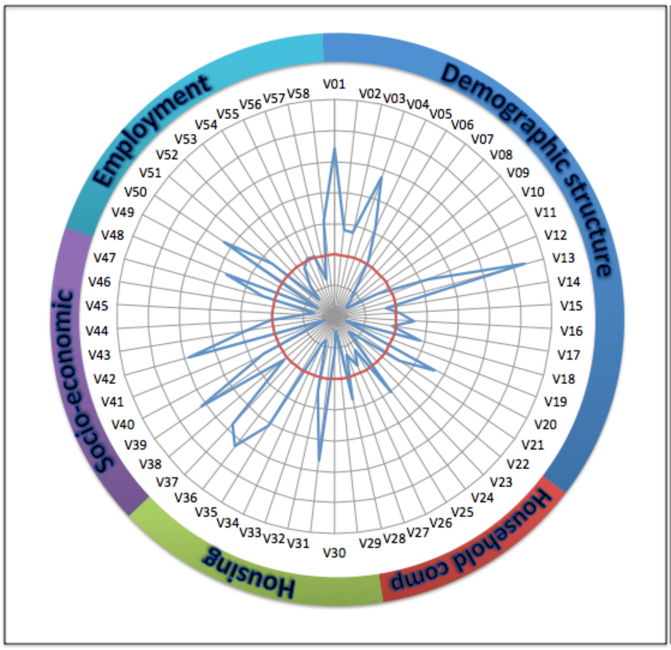
- A radar chart showing, for each variable, the ratio of the mean for JOAs within the cluster to the mean for all JOAs (the red circle denotes a value of 1). Note that these are the variables used in the cluster analysis ie (except for V03, V28 and V38) rates, with an inverse hyperbolic sine transformation, rather than the original census data. This is a simple way of visualising which variables show the greatest divergence from the mean.
- A very crude age profile (by sex) of the total Jewish population within the cluster. The crudeness is due to age breakdowns only being available for 10- or 15-year age bands at the OA level.

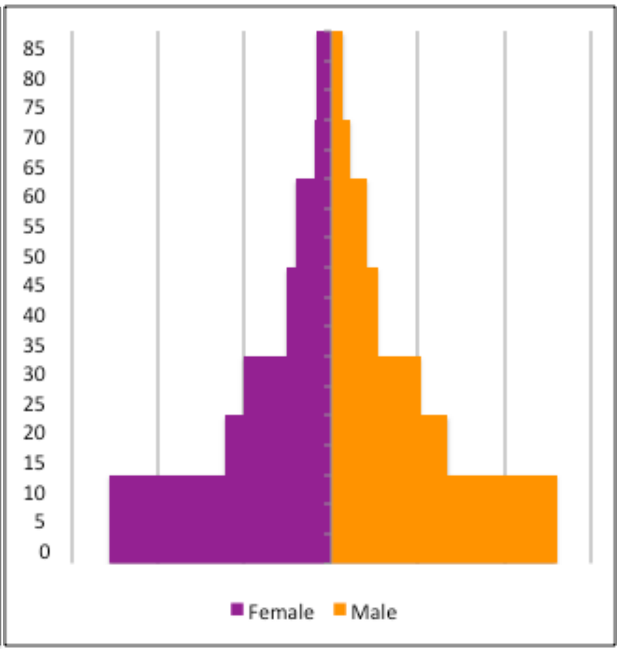
The profiles are followed by a number of illustrative maps. (See References section for copyright information.)

Table 3: Cluster summary statistics

Cluster	A	B	C	D	E	F	G	H	Total
Total People	66,219	69,050	618,976	7,821,098	3,033,704	959,809	29,703,398	13,803,658	**56,075,912**
Total Jews	16,014	18,181	4,517	54,219	35,235	57,003	36,734	41,443	**263,346**
Jews %	24.2%	26.3%	0.7%	0.7%	1.2%	5.9%	0.1%	0.3%	**0.50%**
Total Area (km^2)	4	10	143	1,897	274	247	135,004	13,426	**151,005**
Number of JOAs	51	58	15	182	118	187	126	138	**875**

CLUSTER A: *HAREDI* (STRICTLY ORTHODOX)





Cluster A has the youngest age profile of any of the clusters, with 48% of the population aged under 16. A higher than average proportion of the population were born in continental Europe, in the Middle East, or in the Americas, and the proportion declaring a non-UK national identity was similarly high. Those in this cluster were much more likely than those in any other cluster to describe their ethnicity as “other”.

The cluster is characterised by the presence of very few one-person or pensioner-only households, and the highest “household size” of any cluster (around 4.4).

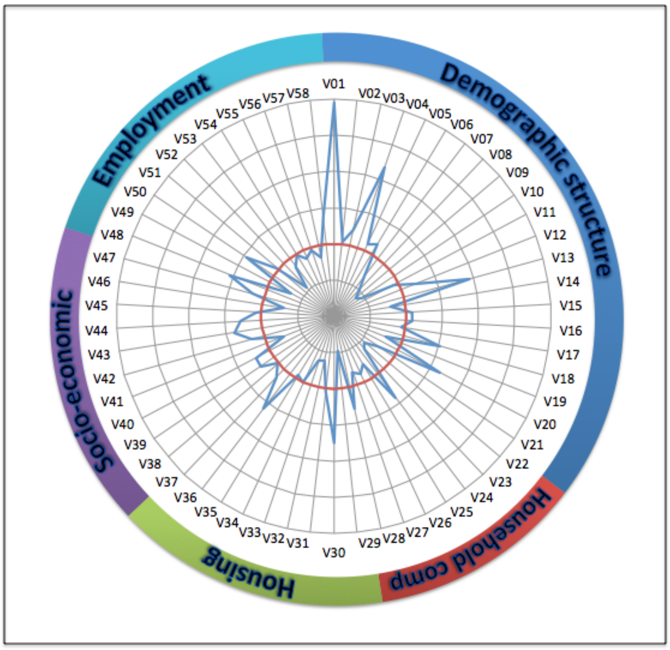
Terraced housing is relatively much more common than in other clusters, as are both private and social renting. Overcrowding is also prevalent, and rates of car ownership are low.

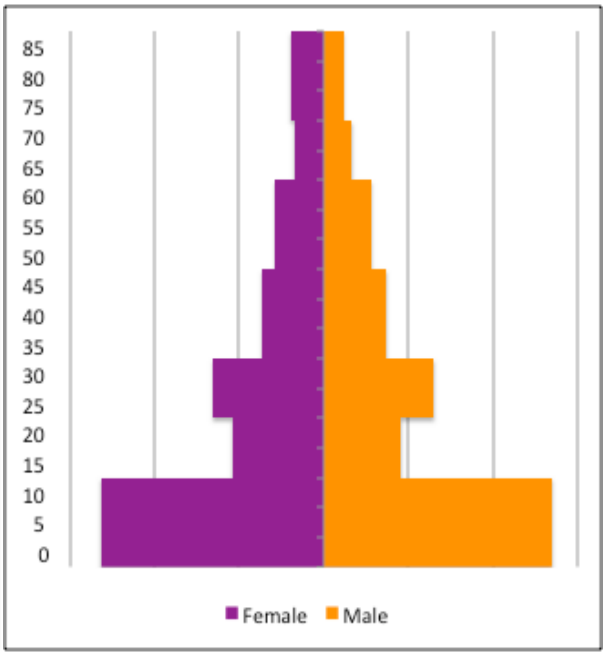
Almost 44% of those aged over 16 reported having no qualifications at all. A high proportion of people were recorded as working part-time, and unemployment rates were also high. Employment within the “public administration, education, health and social work” grouping was above average, but other white-collar occupations were under-represented.

Geographically, 46 of the 51 JOAs in Cluster A are located in Stamford Hill (and cover most of that area); 2 are in Broughton Park in Manchester, and 3 are in the Bensham district of Gateshead. The total area covered by Cluster A is only 4 km^2, the population density is very high, and almost a quarter of the entire population is Jewish. 11.5% of the total (ie Jewish and non-Jewish) population in Cluster A chose not to answer the religion question, compared to 7.2% nationally.

All the JOAs in Cluster A are thus located in areas which would normally be considered to be *haredi* (strictly Orthodox), with very insignificant populations of non-*haredi* Jews. The characteristics cited above are all consistent with this as well – in particular, large families with very low levels of secular education and a large proportion of the population (probably predominantly female) involved in teaching and the caring professions.

CLUSTER B: *HAREDI*-LITE





Cluster B could perhaps best be described as "Cluster A, but less so". The age profile again is very young – 44% under 16. Similarly high proportions of the population were born in continental Europe or the Middle East, but few than average in the Americas. Again, a very high proportion chose to identify ethnically as "other".

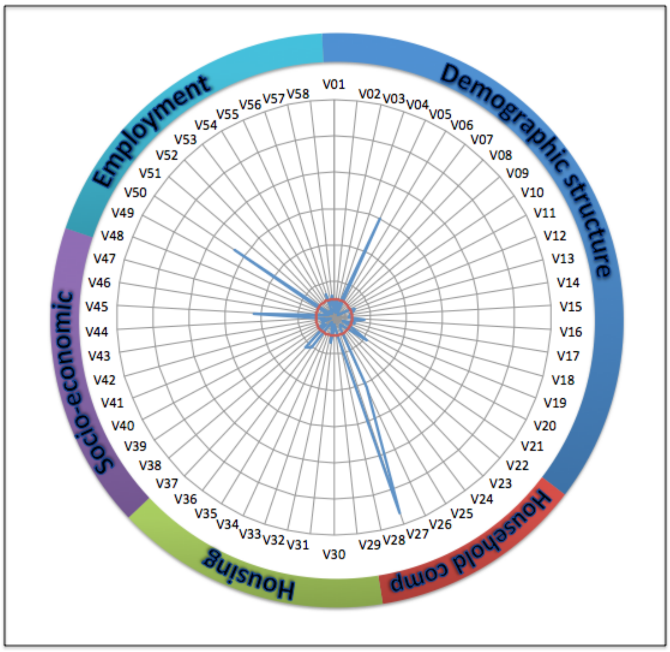
"Household size", at 3.9, is not much below that of Cluster A, but semi-detached housing predominates rather than terraced houses. The proportion renting privately is still much higher than average.

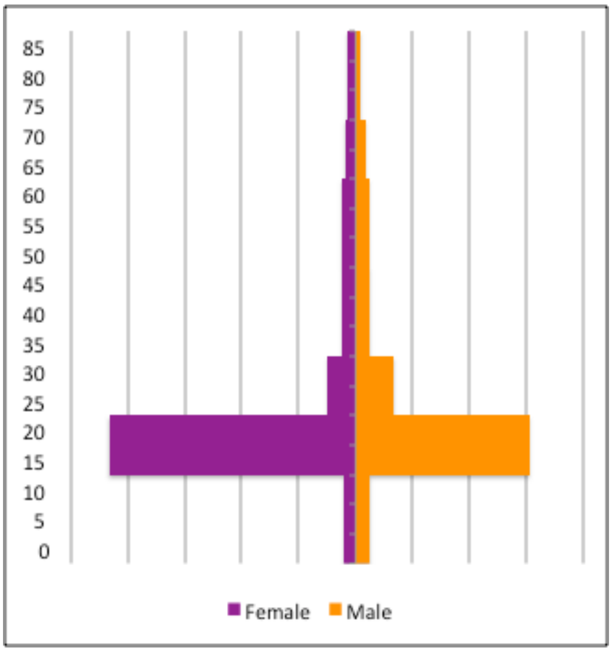
Qualification levels are also below average, with 51% of those over 16 having acquired fewer than 2 "A"-Levels or equivalent, compared to 40% nationally. Employment patterns are similar to Cluster A.

JOAs in Cluster B are found adjacent to those in Cluster A in the Stamford Hill (London) and Bensham (Gateshead) areas, but around half of the Jews in Cluster B are in the London Borough of Barnet – specifically in the Golders Green, Hendon and Edgware neighbourhoods. In Manchester, too, whilst a significant number of Cluster B JOAs are located in Broughton Park, there is also a significant agglomeration around a kilometre to the north, in the Sedgley Park area of the Borough of Bury.

The geographical distribution of Cluster B, and the similarities in characteristics with Cluster A, suggest strongly that Cluster B also represents a segment of the *haredi* community, but a less "extreme" part of it. It is also likely, however (based both on the evidence from the data, and on the author's knowledge of these areas), that those Cluster B JOAs in Barnet and in Sedgley Park include some non-*haredi* (but nevertheless religiously observant) Jews.

CLUSTER C: STUDENTS





Cluster C is the smallest cluster (with a total Jewish population of only 4,500), and also the easiest cluster to define. 65% of the Jewish population are aged between 16 and 24, and 33% live in communal establishments (a category which includes retirement homes but, in this context, is almost certainly student accommodation).

Other characteristics (high proportions of private renting and overcrowding amongst those not in communal establishments, almost 80% with 2 or more "A"-Levels or higher qualifications and, most obviously, 70% identifying as students) are consistent with the identification of Cluster C as student areas.

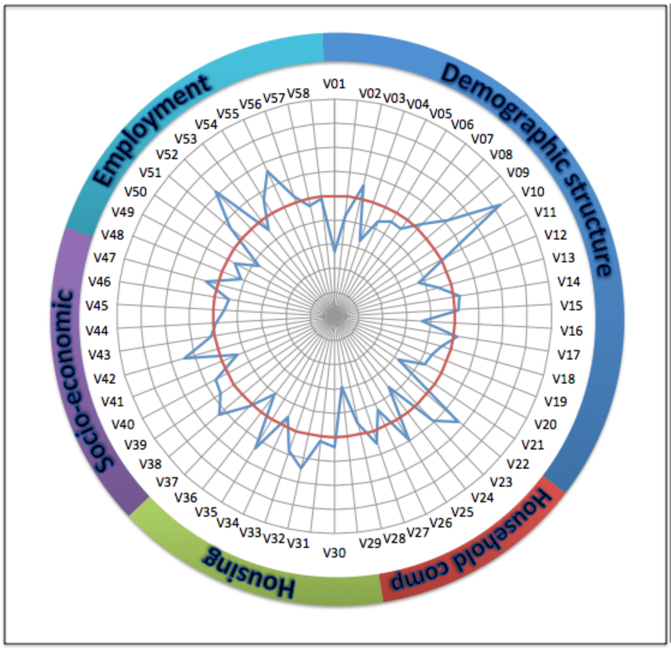
3.2% of Jews in Cluster C were born in the Americas (compared to 1.7% of the Jewish population as a whole), but only 1.1% in Africa (compared to 1.6% overall). Unfortunately it is not possible to split out, from the African-born, what are likely to be the two largest components – those born in the Maghreb and Egypt (who would be identified as part of the "Sephardi" Jewish community), and those from South Africa.

There are only 15 JOAs in Cluster C, of which 1 is in Bristol, with 2 each in Oxford, Cambridge, Birmingham and Leeds, and 3 in Nottingham. In Leeds, which has the largest Jewish population in the UK after London and Manchester, it is very noticeable that the student areas are located some considerable distance from the main Jewish population centre.

Two of the remaining Cluster C JOAs are in Greater Manchester. One of these covers the main student area of Fallowfield in the City of Manchester; the other lies in the City of Salford, between Broughton Park and Sedgley Park. Of the 335 Jews in this particular JOA, 149 are females aged 16 to 24, and 144 live in communal establishments. A little bit of investigation reveals that this JOA includes a major girls' religious seminary.

The last Cluster C JOA straddles Gateshead and Newcastle-upon-Tyne, but examination of the constituent MSOAs reveals that the majority of the Jewish population are on the Gateshead side of the River Tyne. Around 80% of the Jewish population are females aged 16 to 24, and a similar proportion live in communal establishments; again, there is a major seminary located here.

CLUSTER D: OLDER COMMUNITIES



As with Clusters A, B and C, age is one of the main defining characteristics of Cluster D. Here, however, the population pyramid is inverted – 31% of the Jewish population are aged 65 or older (almost 20% are 75 or older). Nevertheless, a majority of the population are still of working age.

As would be expected from the age profile, Cluster D has over-representation of single-person households and of households comprising only people aged 65 and over, and of those who are retired. More surprisingly, "lone parent" families are also over-represented.

A high proportion have no qualifications (reflecting, perhaps, lower levels of participation in formal education in previous generations). Those in Cluster D are more likely than those in any other cluster to work in (or have previously worked in) routine or semi-routine occupations. They are also the most likely to own their home without a mortgage.

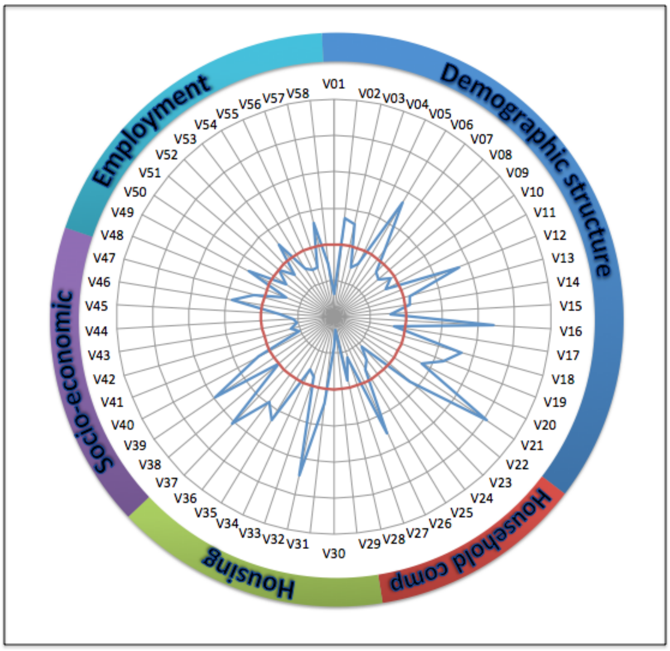
Pockets of Cluster D appear in many parts of the country, almost all of them in areas with organised Jewish communities of a reasonable size – Newcastle-upon-Tyne in the North East region; Manchester (excluding the *haredi* areas), Liverpool and Blackpool / Lytham St Annes in the North West; Leeds and Sheffield in Yorkshire & The Humber region; Birmingham in the West Midlands; Bournemouth / Poole in the South West; Brighton &

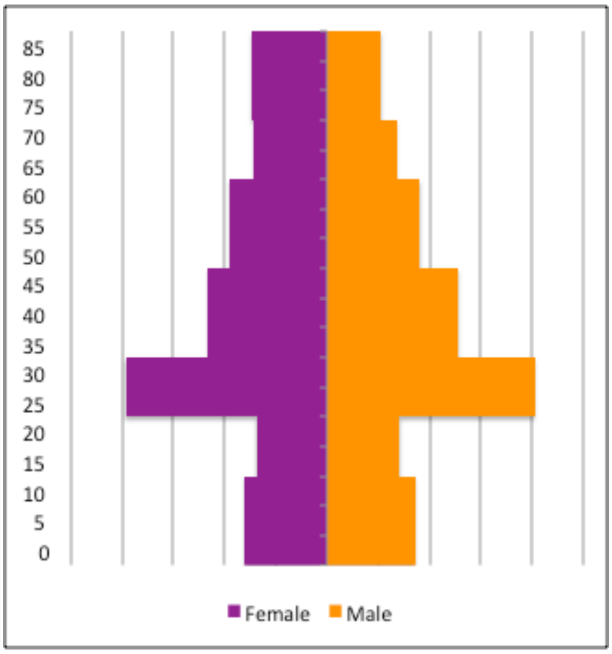
Hove / Worthing in the South East; and Southend-on-Sea in the East of England. Perhaps surprisingly, Portsmouth appears in this cluster whereas nearby Southampton, often thought to be similar in nature, does not.

Most of the outer area of Greater London (by land area, although only a quarter of the Jewish population) is also classified as Cluster D, together with small areas of neighbouring Bushey and Borehamwood.

Cluster D includes many areas which had once been major centres of Jewish population, but which showed significant declines between the 2001 and 2011 censuses – amongst them, most of the provincial communities noted above, and the London Boroughs of Brent, Ealing, Enfield and Redbridge. The name given to this cluster – "Older Communities" – reflects both the age profile, and the reducing significance of the communities located here.

CLUSTER E: YOUNG COSMOPOLITAN PROFESSIONALS





Those aged 25 to 34 are over-represented in Cluster E. Just over a third were born outside the UK (including 10% in continental Europe, 9% in the Americas, 8% in the Middle East or Asia) – by far the highest proportion of any cluster.

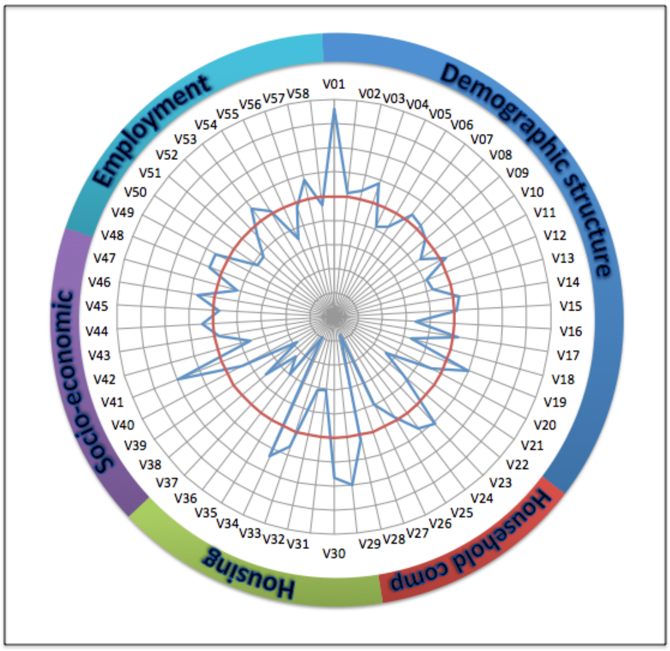
One-person households are very common, as are “other” household types, which would include instances of two or more people living together who are unrelated (and not in a relationship with each other). More than 60% of Jews in Cluster E live in flats, and rates of renting (both private and social) are very high. Many live in accommodation which would be defined as overcrowded.

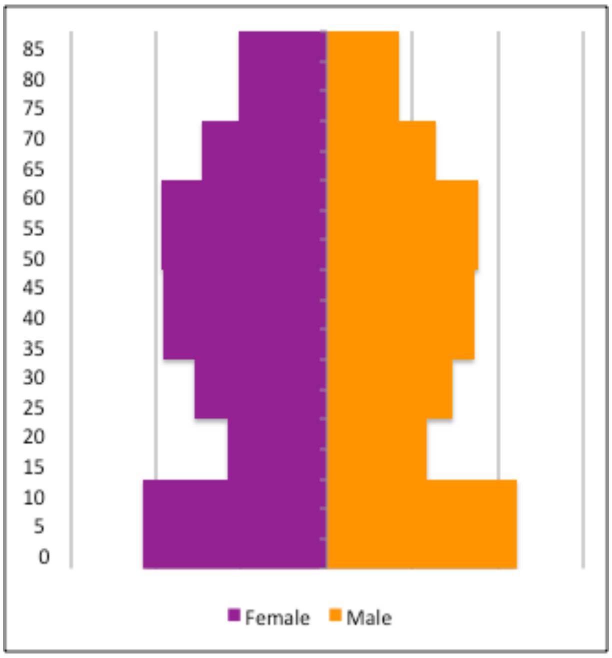
People in Cluster E are more highly qualified on average than any other cluster, with 60% of those aged 16 and over having a degree, compared to just 42% among the Jewish population as a whole. They are also the cluster most likely to be in full-time employment, to be in managerial, administrative or professional roles, and to work in occupations such as finance and property.

Almost all of the JOAs in Cluster E are in central London, with the greatest concentration around West Hampstead and St John’s Wood, although there are a few pockets in outer North West London, and three clusters in Brighton & Hove.

Many of these areas are not well served by synagogues or other Jewish communal facilities, although there are exceptions. It might perhaps be hypothesised that many Jews living in these locations are concentrating on establishing their careers, and are less interested in Jewish life for now, although that might change once they start raising families.

CLUSTER F: ENGAGED FAMILIES





This is numerically the largest cluster, accounting for over one in five of all Jews in England and Wales (just ahead of the Older Communities of Cluster D) and, apart from the two *haredi* clusters (A and B) shows the highest concentration of Jews, at 5.9% of the population.

Those aged under 16 are also more over-represented here than in any clusters apart from A and B. Those aged 16 to 24 are under-represented, which may be explained by many of them having been away at University at the time of the census. Families with dependent children are over-represented among household types.

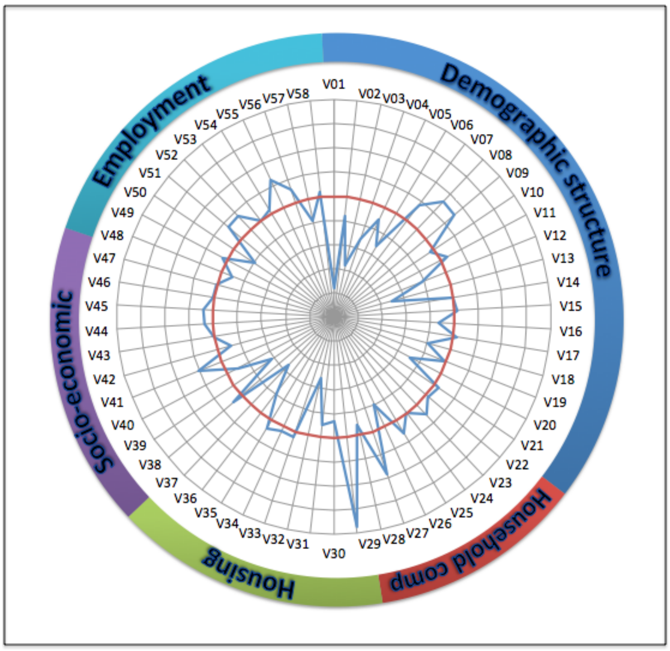
Detached and semi-detached houses predominate in Cluster F, with disproportionately high rates of ownership (whether with or without a mortgage), and of households with two or more cars.

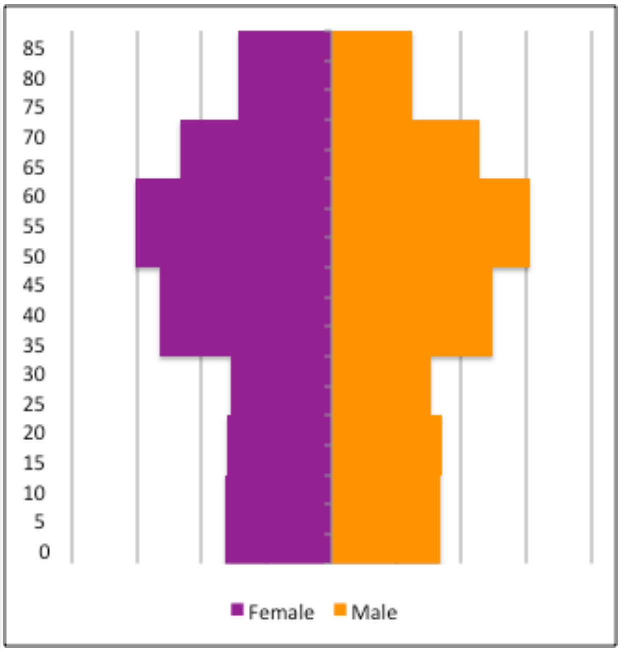
Both part-time and full-time working are well represented, and unemployment rates are the lowest (except for the Students of Cluster C, many of whom would presumably not be looking for work in any case). As with Cluster E, professional and administrative occupations, and the finance and property sectors, are over-represented.

Geographically, all JOAs in Cluster F are located in or very close to areas with established Jewish community organisations. Between them, the North West London Boroughs of Barnet and Harrow, together with the neighbouring Borough of Hertsmere just over the Hertfordshire border, account for just short of two-thirds of all Jews in Cluster F. There is a similar but much smaller concentration around the border of the North East London Borough of Redbridge with Epping Forest in Essex. In South London, Cluster F JOAs are present in Bromley, Kingston-upon-Thames / Sutton and Richmond-upon-Thames, and further out in the commuter belt, in St Albans (Hertfordshire) and Southend-on-Sea (Essex).

In Greater Manchester, elements of Cluster F are found in Broughton Park, Sedgley Park and Whitefield in the north, and Cheadle and Hale in the south. In Leeds, they are found in the Alwoodley area; and finally, there is one JOA in Hove.

CLUSTER G: DISCONNECTED



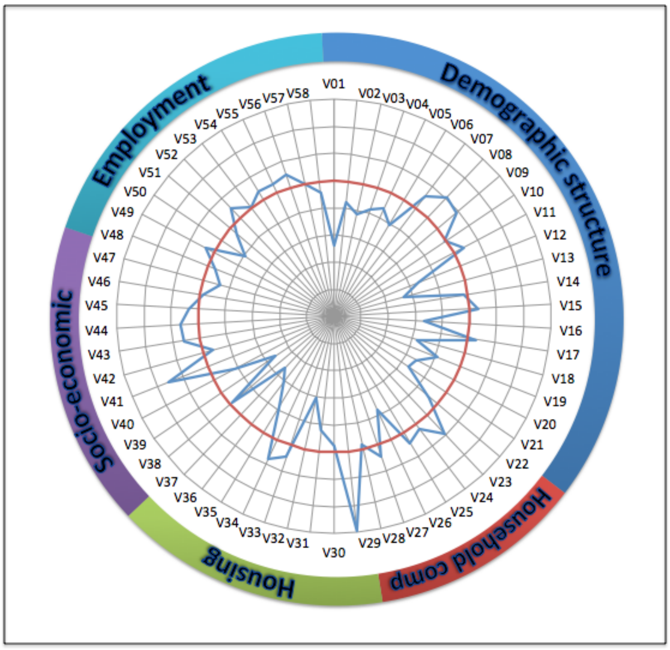


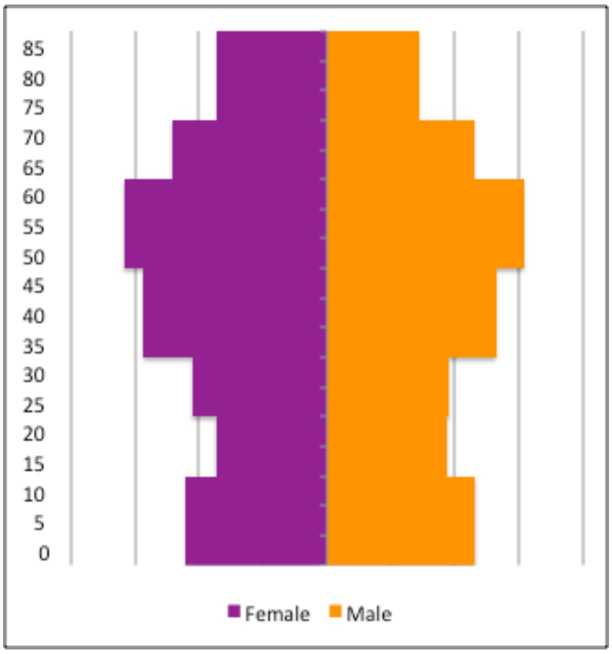
Cluster G covers close to 90% of the land area of England & Wales, and includes 53% of the total population, but only 14% of the Jewish population, and Jews constitute on average only just over 0.1% of the population in these JOAs.

It has one of the older age profiles, with the 50 to 64 age group being the largest. A higher than average percentage (almost 4%) live in communal establishments, but more than 40% live in detached properties. This is perhaps not surprising for the only section of the Jewish population which is not heavily concentrated in major urban areas. Those Jews in Cluster G were least likely to define their ethnicity as "Other", and proportionately fewer than in most other clusters were born outside the UK.

Many of the Jews leaving in these JOAs will inevitably be located a long distance from their nearest Jewish communal facilities; this cluster can therefore be characterised as being somewhat disconnected from the Jewish community.

CLUSTER H: ON THE MARGINS





Cluster H is very similar to Cluster G, although with a slightly younger age profile. It commonly appears in areas surrounding some of the main Jewish population centres, characterised by Clusters D, E and/or F (Greater London, extending along both the Thames Valley to the west and the Thames Estuary to the east; Manchester; Leeds; Liverpool; Birmingham; Brighton & Hove / Worthing; Bournemouth; Newcastle / Gateshead), as well as some student centres (Oxford, Bristol, Nottingham). Cluster H also appears in isolation in Eastbourne, Leicester, Northampton and South Wales.

In effect, Cluster H acts as a margin between Cluster G and the more engaged clusters.

Figure 4: All Clusters in England & Wales (Regions Outlined)

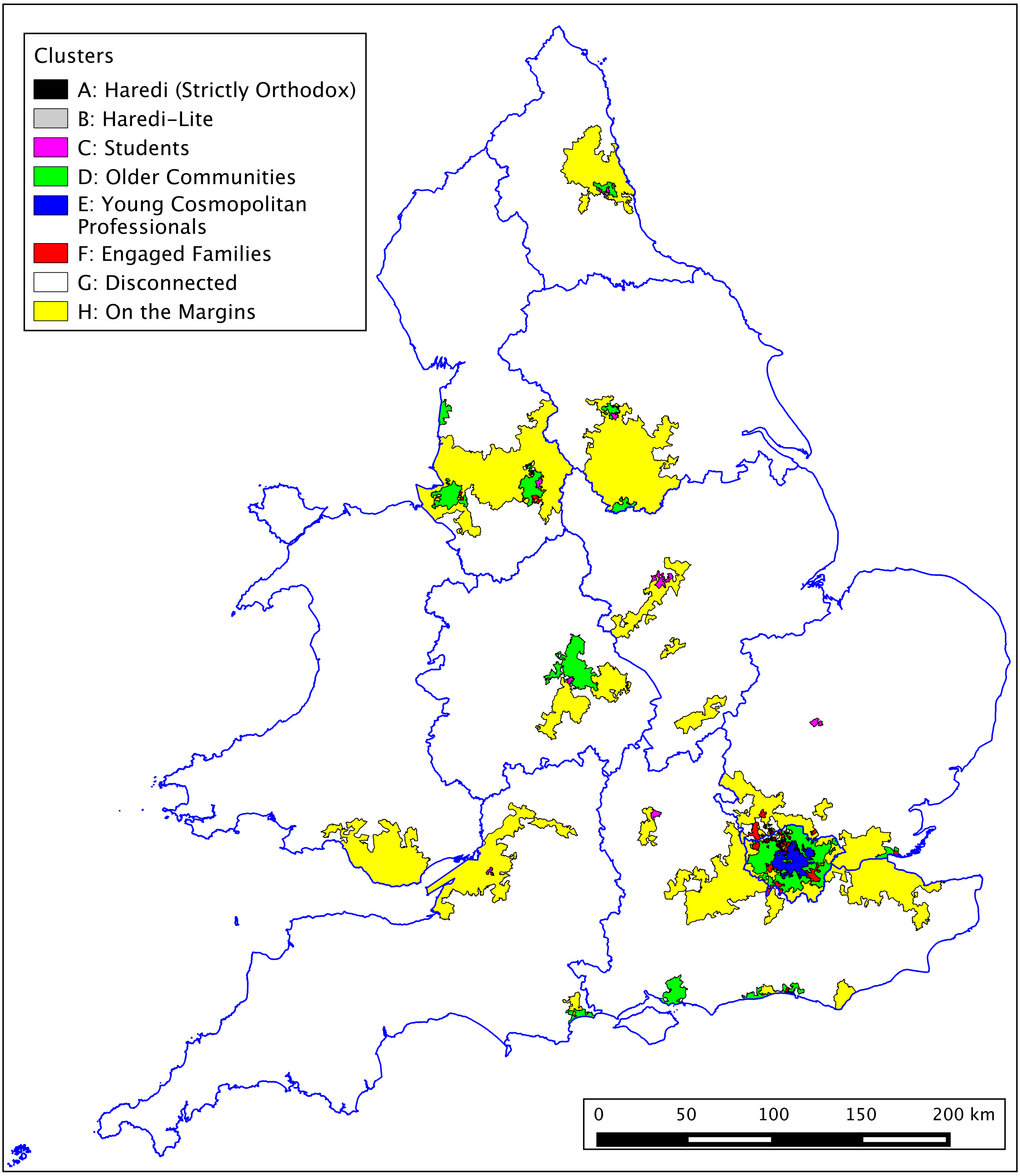


Figure 5: Greater London

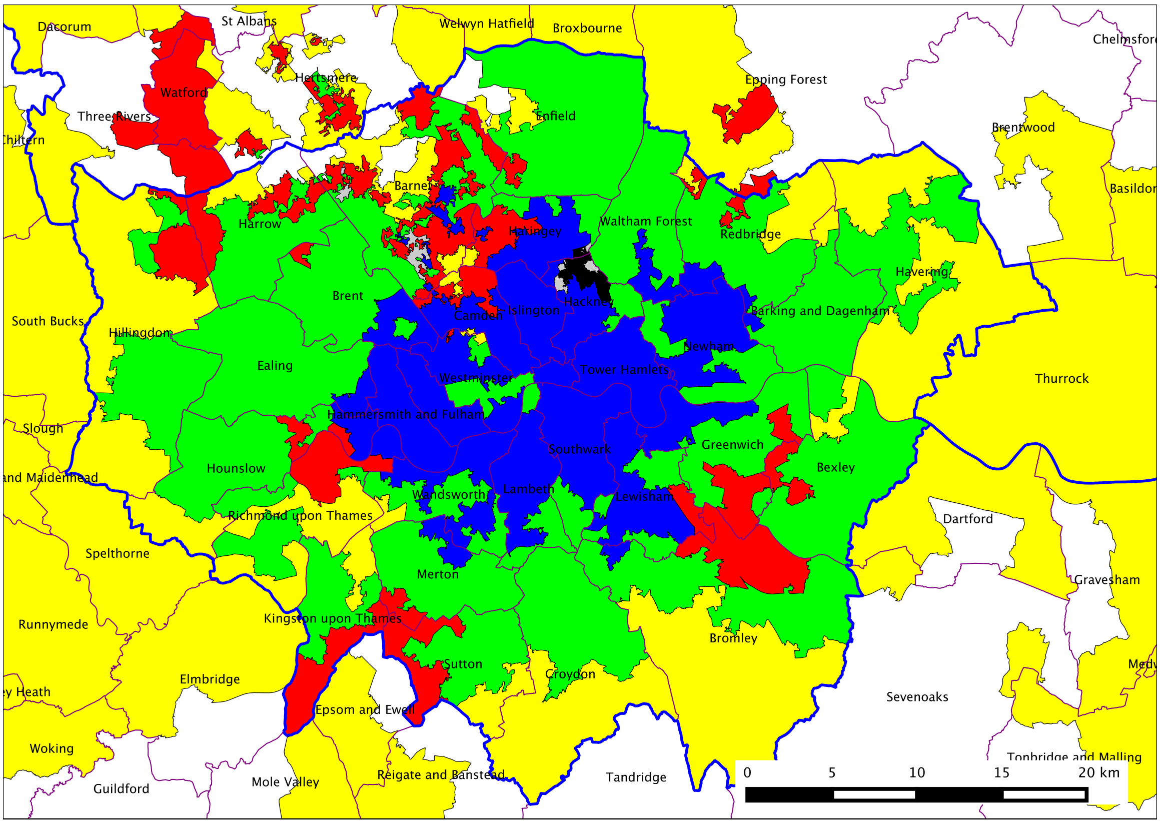


Figure 6: LB Barnet (south)

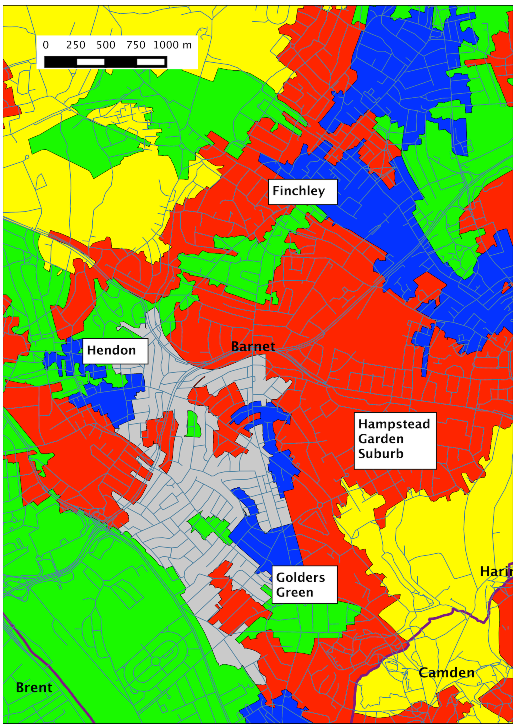


Figure 7: Stamford Hill

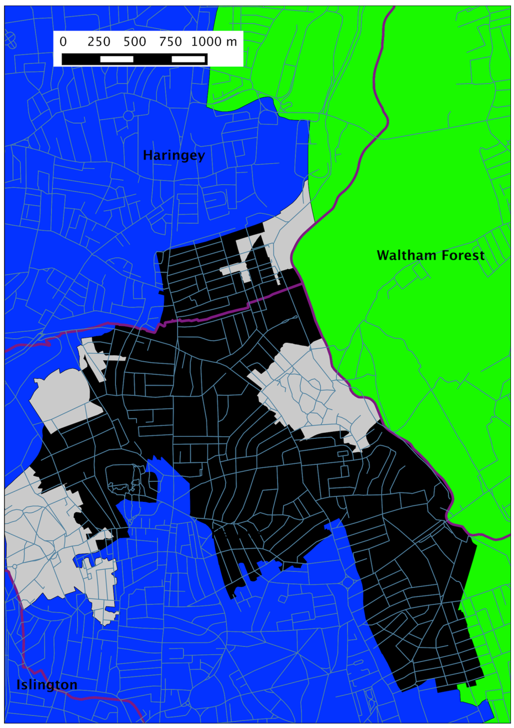


Figure 8: North Manchester

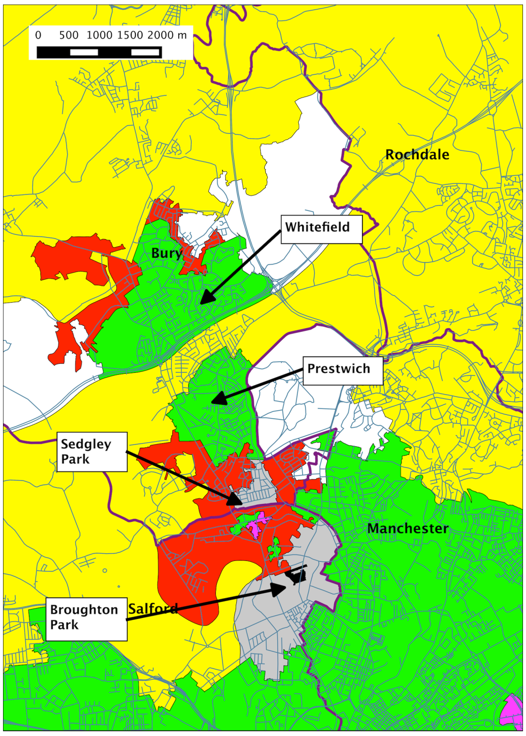


Figure 9: Leeds

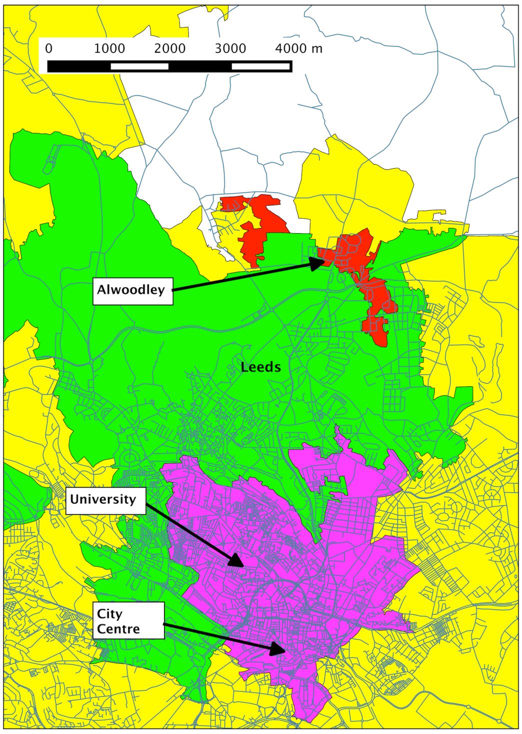


Figure 10: Gateshead / Newcastle-upon-Tyne

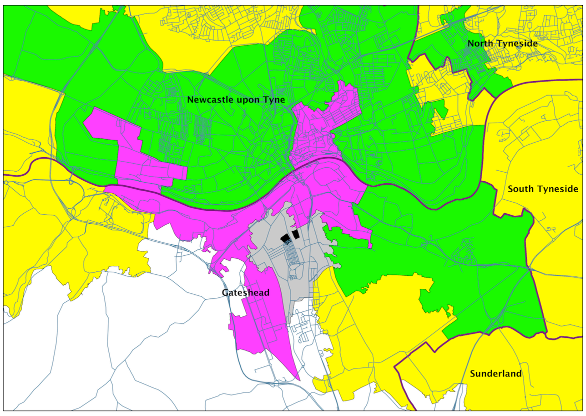
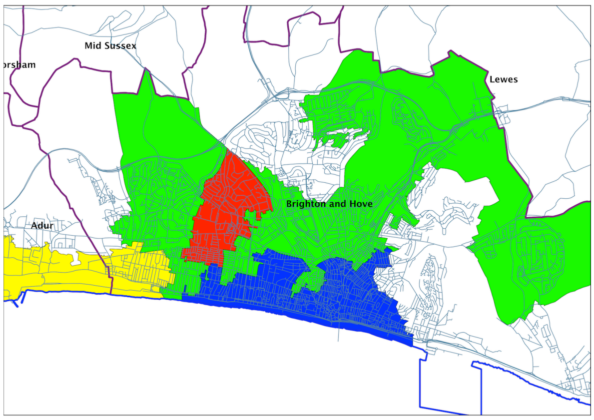


Figure 11: Brighton & Hove



CONCLUSION

The primary objective of the research – carrying out an area classification of the Jewish data from the 2011 Census – was achieved. A set of eight clusters was derived using a k-means clustering algorithm, differentiating characteristics of the clusters were identified, and possible interpretations were attached to the clusters. In some cases (such as the Strictly Orthodox of Cluster A, and the students of Cluster C), these were very obvious; in others, the identifications may be a bit more tentative.

It should be recalled here that the identification of a distinct clustering pattern for Stamford Hill was one of the constraints imposed when selecting the optimal solution – it would have been very surprising if the characteristics of Cluster A had not met expectations. Nevertheless, it was pleasing to see that Cluster A and Cluster B ("*Haredi*-Lite") did between them pick up almost all the areas of known strictly Orthodox concentration, and there were no false positives (ie areas identified as falling in Cluster A or Cluster B which, based on the author's knowledge, clearly shouldn't have).

The appearance of some JOAs in Cluster B (but not Cluster A) in the Hendon and Golders Green areas of North West London (see Figure 3) was also welcome, lending support to the widely believed hypothesis that the strictly Orthodox Jews of North West London are "not quite as *haredi*" as those in Stamford Hill.

The classification of the area between Broughton Park and Sedgley Park in Manchester (see Figure 5) mainly into Cluster F (Engaged Families), with some C (Students) and D (Older Communities), rather than into A and B, was slightly surprising. It remains to be seen whether this is anomalous, or whether it does reflect facts on the ground.

The presence of single Cluster C JOAs in Broughton Park (Figure 5) and Gateshead (Figure 7) pointed to the existence in these locations of seminaries.

It became very clear during the course of the project that there could be no perfect solution. The selection of variables, the weights applied (explicitly, or implicitly through choices around standardisation), and the randomness intrinsic to the k-means algorithm, mean that there could be many solutions which provide reasonable explanations, and it would be harsh to blame an investigator for making the additional suppositions that might be necessary to support such an explanation. Further, since every JOA has to be allocated to exactly one cluster, there is no room for nuance; JOAs which could potentially fit into more than one cluster have to be arbitrarily allocated to one, with no means of flagging up their marginality.

With hindsight, a different approach to the creation of the JOAs may have been adopted. The only constraints applied were that JOAs should consist of contiguous building blocks, should not cross boundaries between regions, and should have a size as close as possible to 300. This led to the creation of a lot of odd-shaped JOAs, some of which would have clearly had poor internal homogeneity.

It would appear that many of the JOAs cover relatively large geographical areas, even though the Jews within them are highly concentrated. This situation would arise due to the need to "mop up" odd Jews in areas of low concentration in order to achieve the mean. One example would be the sole Cluster C JOA in Gateshead / Newcastle-upon-Tyne (Figure 7), which straddles the two towns and includes a couple of individual MSOAs with very few Jews. A more homogenous arrangement would have kept the River Tyne as a boundary, separating the two very different communities here.

An approach which might have achieved this would have been to impose additional constraints based on variables other than the number of Jews. This would of course involve a subjective decision of which variables are most likely to represent homogeneity; but population density would have been a very obvious one to adopt, in order to avoid the presence of urban JOAs with large but sparsely populated rural hinterlands.

Less weight could also have been attached to keeping the JOAs roughly of the same size; a variety of sizes (within reason) should not have adversely affected the method.

With more time available, it is possible that the analysis could be improved by an iterative process, of identifying what appear to be the most important factors, and revising the weight attached to those variables accordingly.

Another possibility would be to attempt a multi-tier analysis (as carried out for the national classifications); an initial attempt to split all JOAs into just two clusters (*ḫaredi* and non-*ḫaredi*) might have been informative.

The author remains confident that an analysis of this type could also be carried out for other religious groups (or subgroups defined by other characteristics such as ethnicity or country of birth). However, the input of someone familiar with that particular community would be invaluable.

[9,202 words, excluding tables, references and appendix]

REFERENCES

Cockings, S., Harfoot, D., Martin, D., Hornby, D., 2011. *Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales*. **Environment and Planning A**, 43(10), 2399-2418. doi:10.1068/a43601

Everitt, D., Landau, S., Leese, M., Stahl, D., 2011. *Cluster Analysis (5th Edition)*. London: Wiley.

Graham, D., Schmool, S., Waterman, S., 2007. *Jews in Britain: A snapshot from the 2001 Census*. London: Institute for Jewish Policy Research. http://www.jpr.org.uk/publication?id=362

Graham, D., Vulkan, D., 2010. Synagogue Membership in the United Kingdom in 2010. London: Institute for Jewish Policy Research / Board of Deputies of British Jews. http://www.jpr.org.uk/publication?id=114

Graham, D., Boyd, J., Vulkan, D., 2012. *2011 Census Results (England and Wales): Initial Insights about the UK Jewish Population*. London: Institute for Jewish Policy Research / Board of Deputies of British Jews. http://www.jpr.org.uk/publication?id=692

Harfoot, A., Cockings, S., Hornby, D., 2010. *Technical summary: Office for National Statistics 2001 Output Area Production System (OAPS) methodology*. School of Geography, University of Southampton.

Holman, C. and Holman, N., 2002. *Torah, worship and acts of loving kindness: baseline indicators for the* charedi *community in Stamford Hill.* Leicester: De Montfort University. http://www.hackney.gov.uk/Assets/Documents/charedi_community_baseline_indicators.pdf

Makles, A., 2012. *Stata tip 110: How to get the optimal k-means cluster solution*. **The Stata Journal**, 12(2), 347-351.

Martin, D., 2003. *Extending the automated zoning procedure to reconcile incompatible zoning systems*. **International Journal of Geographical Information Science** 17 181-196.

Martin, D., Nolan, A. and Tranmer, M., 2001. *The application of zone-design methodology in the 2001 UK Census*. **Environment and Planning A,** 33(11), 1949-1962.

Office for National Statistics (ONS), 2014. *Methodology Note for the 2011 Area Classification for Output Areas*. http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/methodology-and-variables/methodology.pdf

Openshaw, S., 1977. *A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling*. **Transactions of the Institute of British Geographers, New Series,** 2, 459-472.

Tait, A., 2012. *Changes to Output Areas and Super Output Areas in England and Wales, 2001 to 2011*. Office for National Statistics.

Vickers, D., Rees, P., Birkin, M., 2005. *Creating the National Classification of Census Output Areas: Data, Methods and Results.* Working Paper, School of Geography, University of Leeds. http://www.geog.leeds.ac.uk/fileadmin/documents/research/csap/05-02.pdf

Vulkan, D. and Graham, D., 2008. *Population Trends among Britain's Strictly Orthodox Jews.* London: Board of Deputies of British Jews. http://www.bod.org.uk/content/StrictlyOrthodox.pdf (also at http://www.ukjewish.info/files/StrictlyOrthodox.pdf)

SOFTWARE APPLICATIONS

The following specialist software tools were utilised:

AZTool / AZTImporter (construction of output areas)

Copyright David Martin, Samantha Cockings and University of Southampton

http://www.geodata.soton.ac.uk/software/AZTool/

See also Cockings et al (2011), Martin (2003).

Stata (cluster analysis)

StataCorp LP

http://www.stata.com/

QGIS (geographical visualisation)

http://www.qgis.org/en/site/

DATA SOURCES

Secondary data were obtained from the following sources:

Office for National Statistics (ONS)

- 2011 census data obtained via Nomis (https://www.nomisweb.co.uk/).

- Boundary data obtained via Open Geography Portal (https://geoportal.statistics.gov.uk/).

OpenStreetMap (http://www.openstreetmap.org/) and **MapCruzin** (http://www.mapcruzin.com/)

- Shapefiles for roads and waterways (note that, due to different projection methods, locations are approximately 110m WNW of the corresponding locations in the ONS boundary data).

APPENDIX

Item 1: AZTool parameter file for generation of JOAs from MSOAs


```
<?xml version="1.0" encoding="utf-8"?>
<ProgramOptions xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <InputPATFile>{path omitted}\MSOA_2011.pat</InputPATFile>
  <InputAATFile>{path omitted}\MSOA_2011.aat</InputAATFile>
  <Header>true</Header>
  <IDIndex>1</IDIndex>
  <RegionIndex>7</RegionIndex>
  <RegionToUse>ALL</RegionToUse>
  <RespectRegions>true</RespectRegions>
  <TargThreshVars>
    <TargetThresholdVar>
      <Name>Jews</Name>
      <FileIndex>4</FileIndex>
      <TargetSet>true</TargetSet>
      <Target>300</Target>
      <Tolerance>1000000</Tolerance>
      <Weight>100</Weight>
      <MinThreshSet>true</MinThreshSet>
      <MinThresh>100</MinThresh>
      <MaxThreshSet>true</MaxThreshSet>
      <MaxThresh>500</MaxThresh>
    </TargetThresholdVar>
  </TargThreshVars>
  <AreaIndex>8</AreaIndex>
  <P2ASet>false</P2ASet>
  <P2AWeight>100</P2AWeight>
  <MinBdyLenSet>false</MinBdyLenSet>
  <MinBdyLenPerc>10</MinBdyLenPerc>
  <IgnoreBishopsContig>true</IgnoreBishopsContig>
  <AllowDonuts>true</AllowDonuts>
  <IRATargetBasedTractCount>true</IRATargetBasedTractCount>
  <TestSpreadsheetReqd>false</TestSpreadsheetReqd>
  <ReportStatistics>true</ReportStatistics>
  <NumberSwapIterations>20</NumberSwapIterations>
  <NumberRuns>20</NumberRuns>
  <UseLogDomainScores>false</UseLogDomainScores>
  <IgnoreTractsWithUnbreachedBB>false</IgnoreTractsWithUnbreachedBB>
  <RandomSeed>0</RandomSeed>
</ProgramOptions>
```

Item 2: Stata script for k-means optimisation

```
* do cluster analysis (1-9 clusters)

forval k = 1/9 {
cluster kmeans V01-V58 , k(`k')
}

* WSS matrix
matrix WSS = J(9,5,.)
matrix colnames WSS = k WSS log(WSS) eta-squared PRE

* WSS for each clustering
forval k = 1/9 {
	scalar ws`k' = 0
	foreach v of varlist V01-V58 {
		quietly anova `v' _clus_`k'
		scalar ws`k' = ws`k' + e(rss)
	}
	matrix WSS[`k', 1] = `k'
	matrix WSS[`k', 2] = ws`k'
	matrix WSS[`k', 3] = log(ws`k')
	matrix WSS[`k', 4] = 1 - ws`k'/WSS[1,2]
	matrix WSS[`k', 5] = (WSS[`k'-1,2] - ws`k')/WSS[`k'-1,2]
}


matrix list WSS

* plot graphs
local squared = char(178)
_matplot WSS, columns(2 1) connect(l) xlabel(#10) name(plot1, replace) nodraw noname
_matplot WSS, columns(3 1) connect(l) xlabel(#10) name(plot2, replace) nodraw noname
_matplot WSS, columns(4 1) connect(l) xlabel(#10) name(plot3, replace) nodraw noname
ytitle({&eta}`squared')
_matplot WSS, columns(5 1) connect(l) xlabel(#10) name(plot4, replace) nodraw noname
graph combine plot1 plot2 plot3 plot4, name(plot1to4, replace)

*tabulate means
tabstat V01-V58, statistics( mean ) by(_clus_8)
```

Item 3: Means of analysed variables by cluster

(See Table 1 for variable definitions)

CLUSTER	A	B	C	D	E	F	G	H	ALL	
Variable										
V01	0.407	0.444	0.053	0.081	0.047	0.259	0.036	0.078	0.150	Demography
V02	0.124	0.092	0.077	0.075	0.122	0.091	0.075	0.075	0.089	Demography
V03	5.834	5.156	4.473	4.662	5.481	4.524	1.837	3.260	4.213	Demography
V04	0.466	0.425	0.060	0.130	0.140	0.228	0.135	0.162	0.196	Demography
V05	0.157	0.110	0.600	0.087	0.080	0.082	0.089	0.087	0.100	Demography
V06	0.128	0.143	0.103	0.113	0.226	0.107	0.084	0.097	0.123	Demography
V07	0.098	0.122	0.065	0.163	0.206	0.191	0.208	0.198	0.179	Demography
V08	0.077	0.095	0.064	0.196	0.157	0.193	0.245	0.225	0.186	Demography
V09	0.025	0.037	0.032	0.118	0.078	0.096	0.125	0.113	0.096	Demography
V10	0.028	0.052	0.031	0.187	0.106	0.097	0.108	0.112	0.113	Demography
V11	0.557	0.659	0.664	0.730	0.578	0.742	0.751	0.766	0.705	Demography
V12	0.244	0.183	0.185	0.111	0.277	0.124	0.116	0.100	0.149	Demography
V13	0.137	0.084	0.045	0.038	0.047	0.036	0.021	0.023	0.043	Demography
V14	0.205	0.192	0.216	0.254	0.260	0.256	0.242	0.236	0.243	Demography
V15	0.542	0.553	0.431	0.529	0.401	0.522	0.525	0.546	0.513	Demography
V16	0.137	0.117	0.182	0.078	0.236	0.072	0.093	0.071	0.107	Demography
V17	0.700	0.705	0.712	0.767	0.624	0.770	0.765	0.787	0.742	Demography
V18	0.077	0.081	0.062	0.049	0.097	0.036	0.043	0.034	0.053	Demography
V19	0.007	0.024	0.018	0.028	0.052	0.038	0.032	0.026	0.032	Demography
V20	0.110	0.099	0.059	0.051	0.082	0.055	0.043	0.041	0.060	Demography
V21	0.044	0.026	0.078	0.022	0.089	0.018	0.037	0.023	0.035	Demography
V22	0.140	0.164	0.306	0.404	0.410	0.226	0.305	0.275	0.299	Household
V23	0.036	0.061	0.052	0.121	0.058	0.124	0.113	0.121	0.102	Household
V24	0.668	0.619	0.255	0.324	0.332	0.509	0.438	0.462	0.441	Household
V25	0.035	0.043	0.037	0.066	0.047	0.053	0.062	0.060	0.055	Household
V26	0.066	0.067	0.322	0.065	0.134	0.060	0.060	0.058	0.076	Household
V27	0.017	0.022	0.307	0.030	0.016	0.004	0.037	0.029	0.027	Household
V28	2.203	2.092	1.754	1.447	1.430	1.691	1.469	1.551	1.609	Household
V29	0.053	0.115	0.091	0.142	0.042	0.337	0.420	0.382	0.240	Housing
V30	0.157	0.469	0.163	0.291	0.112	0.358	0.231	0.255	0.269	Housing
V31	0.415	0.185	0.249	0.185	0.213	0.109	0.160	0.150	0.178	Housing
V32	0.319	0.160	0.167	0.330	0.574	0.160	0.132	0.158	0.256	Housing
V33	0.125	0.201	0.113	0.364	0.273	0.351	0.326	0.339	0.310	Housing
V34	0.178	0.380	0.109	0.350	0.303	0.483	0.397	0.440	0.381	Housing
V35	0.117	0.054	0.023	0.072	0.094	0.011	0.061	0.045	0.057	Housing
V36	0.525	0.319	0.411	0.162	0.296	0.120	0.158	0.120	0.200	Housing
V37	0.193	0.094	0.187	0.083	0.165	0.037	0.044	0.042	0.081	Housing
V38	0.883	0.833	0.459	1.036	0.818	0.608	0.899	0.815	0.828	Sociodem.
V39	0.421	0.199	0.150	0.184	0.322	0.060	0.101	0.085	0.163	Sociodem.

CLUSTER	A	B	C	D	E	F	G	H	ALL	
V40	0.459	0.436	0.226	0.397	0.418	0.307	0.320	0.294	0.357	Sociodem.
V41	0.068	0.312	0.292	0.364	0.219	0.586	0.508	0.552	0.420	Sociodem.
V42	0.433	0.211	0.030	0.230	0.112	0.129	0.159	0.166	0.179	Sociodem.
V43	0.146	0.140	0.073	0.105	0.052	0.098	0.102	0.113	0.101	Sociodem.
V44	0.129	0.163	0.065	0.124	0.071	0.137	0.135	0.143	0.126	Sociodem.
V45	0.060	0.116	0.473	0.098	0.081	0.102	0.117	0.115	0.107	Sociodem.
V46	0.133	0.279	0.279	0.355	0.572	0.457	0.405	0.390	0.400	Sociodem.
V47	0.082	0.079	0.039	0.074	0.078	0.056	0.068	0.058	0.068	Sociodem.
V48	0.351	0.293	0.063	0.156	0.135	0.209	0.170	0.178	0.189	Employment
V49	0.110	0.250	0.151	0.344	0.490	0.407	0.385	0.394	0.368	Employment
V50	0.208	0.141	0.645	0.073	0.077	0.075	0.077	0.076	0.097	Employment
V51	0.030	0.024	0.009	0.026	0.023	0.018	0.026	0.021	0.023	Employment
V52	0.077	0.114	0.057	0.299	0.167	0.193	0.241	0.230	0.210	Employment
V53	0.277	0.385	0.185	0.402	0.558	0.505	0.447	0.451	0.447	Employment
V54	0.259	0.281	0.072	0.322	0.220	0.292	0.296	0.305	0.285	Employment
V55	0.100	0.085	0.034	0.129	0.070	0.069	0.121	0.104	0.097	Employment
V56	0.174	0.164	0.234	0.181	0.122	0.164	0.191	0.188	0.171	Employment
V57	0.189	0.252	0.214	0.284	0.399	0.350	0.248	0.301	0.302	Employment
V58	0.441	0.392	0.344	0.278	0.209	0.264	0.293	0.259	0.283	Employment