Computers, Environment and Urban Systems 56 (2016) 36–47

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Beyond the strictly orthodox/mainstream divide: Applying geodemographic analysis to a small nationwide sub-population

Philip Sapiro *

Department of Geography and Planning, University of Liverpool, Liverpool, UK

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form 9 November 2015
Accepted 16 November 2015
Available online 30 November 2015

Keywords:
Geodemographic
Census
Cluster analysis
Jewish
England and Wales
Minorities

ABSTRACT

The use of geodemographic analysis has a long history, arguably stretching back to Charles Booth's *Descriptive Map of London's Poverty*, produced in 1886 and the published classification of areas has invariably been based on all residents. The work described in this paper, however, is novel in the use of geodemographic analysis to focus on a single minority group within a national census. This paper describes the development of a methodology which allows geodemographic analysis to be applied to unevenly distributed minority sub-populations, overcoming two particular issues: finding a suitable geographic base to ensure data reliability; and developing a methodology to avoid known weaknesses in certain clustering techniques, specifically distortion caused by outlier cases and generation of sub-optimal local minimum solutions. The approach, which includes a visual element to final classification selection, has then been applied to establish the degree to which the Jewish population in an area is similar in character to, or differs from, Jews living in other areas of England and Wales, using data from the 2011 census. That group has been selected because of the maturity of its presence in Britain — study of this group may point the way for examination of other, more recently arrived, sub-populations. Previous studies have generally assumed homogeneity amongst 'mainstream' Jews and have not considered spatial variation, separating out only strictly orthodox enclaves. This paper demonstrates that there are indeed distinct socio-economic and demographic differences between Jewish groups in different areas, not fully attributable to the underlying mainstream social geography, whilst also identifying a strong degree of spatial clustering; it also establishes the practicality of applying geodemographic analysis to minority groups.

1. Introduction

A range of techniques with the aim of subdividing a set of objects into a series of broadly homogenous sub-groups falls under the generic title of 'cluster analysis', sometimes referred to more formally as 'numerical taxonomy' (Everitt, Landau, Leese, & Stahl, 2011; Lorr, 1983). The addition of a spatial element to the analysis differentiates geodemographics from other forms of cluster analysis, and determines whether there is any locational relationship between similarly classified areas. Put perhaps too simply, geodemographics is the 'analysis of people by where they live' (Sleight, 1997, p16).

Geodemographic analysis has an extended history (Batey & Brown, 1995; Singleton, 2004; Singleton & Spielman, 2014) and some geodemographic investigations have included a religion or ethnicity variable in studies of the whole population, such as the study on ethnicity and school choice in Birmingham (Harris, Johnston, & Burgess, 2007). However, the work described in this paper is unusual in the use of geodemographic analysis to classify areas solely on the basis of the characteristics of a single minority group (Jews) within a national census.

So, what is the wider benefit of this research? As with any form of neighbourhood classification, the outputs can be used to identify the needs of the targeted group (for example, assessing future social and community requirements), and previous work in examining socio-economic/demographic issues for Jews in England and Wales has only been able to make use of geographically limited surveys (for example, Kosmin & Levy, 1981), or small sample national studies (Graham, Staetsky, & Boyd, 2014; also Kotler-Berkowitz, 2006, and Goldstein, 2013 for equivalent American experience). The approach adopted here could be applied to other sub-populations, so why select the Jewish group for this study? The majority of Jews in the UK have their roots in the major migration westwards from the Russian Empire which took place between 1880 and 1914; some chose the UK as their preferred destination, others had hoped or intended to continue on to the USA, but either could not face or afford the second stage of the journey (Endelman, 2002). During the first half of the twentieth century, Jews formed the only significant non-western-European and non-Christian minority group in Britain. Immigration from a range of world regions during the second half of the century now means that Britain is home to overseas and first and second generation UK-born citizens with a

* Dept of Geography and Planning, University of Liverpool, Roxby Building, Liverpool L69 7ZT, United Kingdom.
E-mail address: psapiro@liv.ac.uk.

http://dx.doi.org/10.1016/j.compenvurbsys.2015.11.004
0198-9715/

range of ethnic group and religious backgrounds (Simpson, 2012). However, the majority of Jews in Britain are now third to fifth generation UK born; so a study focused on this well-established group may provide pointers for other groups whose UK presence is less mature (Waterman & Kosmin, 1987).

Previous studies have identified concentrations of strictly orthodox Jews (Vulkan & Graham, 2008; Graham, 2013 in the UK; and Comenetz, 2006 in the USA) and have considered their socio-economic characteristics finding large family sizes and high levels of deprivation (Holman & Holman, 2002; Valins, 2003). Other studies have given some limited attention to spatial variation in the characteristics of 'mainstream' Jews (Abramson, Graham, & Boyd, 2011; Becher, Waterman, Kosmin, & Thomson, 2002; Graham et al., 2014); in most studies, however, this group, whose overall characteristics (as measured by the census) are not dissimilar to the wider UK population, tend to be considered as a homogenous group. This paper develops a methodology to overcome the challenges in applying geodemographic analysis to unevenly distributed minority groups, and applies that approach to establish the degree to which the Jewish population in one area is similar to or differs from Jews living in other areas.

2. Classification techniques and previous census analyses

The data on which clustering techniques are to be applied can usually be presented as an *N* row by *k* column matrix, where each row represents a case or entity, and each column represents one of the characteristics or variables of the cases. The background to and overall process undertaken in cluster analysis and geodemographics are now well established and do not need to be detailed here. Both Lorr (1983) and Everitt et al. (2011) provide comprehensive overviews of the clustering concept; Harris, Sleight, and Webber (2005) provide a briefer synopsis. They note that there are two basic 'families' of clustering techniques: hierarchical (where cases are progressively grouped into clusters) and optimising techniques, which generally commence by subdividing the totality of entities into a number of clusters and then iteratively attempt to improve the clustering by moving the boundaries between clusters.

Analyses based on USA, UK, and other census data (usually in combination with other information) have been carried out by commercial organisations for use primarily as a tool to target marketing campaigns for private-sector organisations (Harris et al., 2005; Singleton & Spielman, 2014; Webber, 1985). Non-commercial analyses of UK census data have been carried out for, or in partnership with, the Office for National Statistics (ONS) from the 1981 census onwards. Local authority level analyses are described by Wallace and Denham (1996), and Vickers, Rees, and Birkin (2003); and analyses based on lower level geographies by Charlton, Openshaw, and Wymer (1985); Blake and Openshaw (1995), and Vickers, Rees, and Birkin (2005). Despite extensive experimentation, in each case the final analyses were produced through the use of Ward's hierarchical approach, or the *k*-means algorithm, or various combinations of the two (see Lorr, 1983, and Everitt et al., 2011, for the detail of these techniques). A multi-level *k*-means assessment was adopted by ONS in carrying out a classification based on the 2011 census (ONS, 2014).

3. Technical challenges

So, what lessons can be learned from previous work in this field? In terms of the techniques, hierarchical methods benefit from a greater transparency in the process and the sequential formation of clusters, but are 'sensitive to outliers' (Everitt et al., 2011, p79; see also Hubert, 1974). In addition, the movement of cluster centres which occurs as clusters are merged can mean that cases close to the periphery of a cluster might be located closer to the centre of a different later-stage cluster — as 'making the best decision at each particular step does not necessarily lead to an optimal overall result' (Harris et al., 2005, p162). Of the hierarchical techniques, Ward's algorithm is the most popular for assessing population census data.

Conversely, the *k*-means approach has the benefit of ensuring that the solution produced does locate every case in the cluster to which it is 'closest'. However, it is something of a 'black box' approach and is highly prone to produce local minimum solutions; as Everitt et al. (2011) indicate, a 100 case, 5 cluster scheme has over 10^{67} possible solutions, and they cannot realistically all be tested. Steinley (2003, 2006) recommends running large *k*-means clustering analyses with at least 5000 different starting points in order to overcome the issue of local minima.

The approach adopted in this paper seeks combine positive elements of both techniques. Whilst combined technique approaches have been used in earlier census-based classifications (Bailey, Charlton, Dollamore, & Fitzpatrick, 2000; ONS, 2003), those assessments used the *k*-means technique solely to re-allocate Ward's algorithm cases to their nearest cluster centre. The approach adopted in the current paper is novel in combining two techniques specifically to address the local minima and outlier issues.

4. Development of the study classification methodology

Prior to carrying out the main assessment which is described in detail in later sections, some extensive preliminary analysis was carried out, using 2011 England and Wales census data for Jewish residents aggregated at local authority level. That analysis confirmed the instability of the *k*-means approach with a variety of solutions (local minima) being produced. It also confirmed that the inclusion or exclusion of outlier cases produced different results over a wide range of number of clusters when using Ward's method, and also using the *k*-means technique.

Bearing in mind everything which has been identified and discussed so far, a hybrid classification approach has been devised, based around:

1 Identifying cases where the closest neighbour distance can be regarded as an outlier.
2 Using Ward's approach to cluster the (non-outlier) cases.
3 Running a *k*-means clustering on the dataset without outliers, using the Ward cluster centres as a starting point, to re-allocate cases to their nearest cluster, and produce final cluster centres.
4 Adding the outliers back into the dataset and, using the final cluster centres, allocating the outliers to classes.
5 Mapping the results and finalising the number of classes to be used.

The methodology avoids using the *k*-means approach with a random starting point — so the issues surrounding the optimisation process do not arise. Similarly, the issue of outliers is taken out of the process; their re-incorporation at the end ensures that all cases can be involved, but outliers do not influence the position of class centres. A preliminary choice of the range of number of clusters can be made early in the process, but this can be revised, and the final choice is left to the qualitative judgement of the researcher so that the purpose to which the classification is to be put can be accounted for (Harris et al., 2005).

5. Development of analysis units (cases)

The modelling approach outlined above may have relevance for many geodemographic analyses. However, attempting a geodemographic classification of a small and unevenly distributed minority group raises a second and more specific challenge: the development of an appropriate geographic base. Output Areas (OAs) are the basic building blocks for census output. OAs were specifically devised by ONS to represent homogenous areas as far as the nature of the residential dwellings contained within them is concerned, and have a typical population of about 300 persons. For analysis purposes, small groups of OAs (typically five) have been linked (by ONS) to form lower layer super output areas (LSOAs), and small groups of LSOAs (again