HISTORICAL LIFE COURSE STUDIES

VOLUME 2

2015

EHPS

NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at http://www.ehps-net.eu/journal.

Editors: Koen Matthijs & Paul Puschmann
Family and Population Studies
KU Leuven, Belgium
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

EUROPEAN
SCIENCE
FOUNDATION

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

EHPS
NETWORK

HISTORICAL LIFE COURSE STUDIES
VOLUME 2 (2015), 38-57, published 29-09-2015

Creating a typology of parishes in England and Wales: Mining 1881 census data

Kevin Schürer
University of Leicester

Tatiana Penkova
Institute of Computational Modelling SB RAS

ABSTRACT

The paper presents the application of principal component analysis and cluster analysis to historical individual level census data in order to explore social and economic variations and patterns in household structure across mid-Victorian England and Wales. Principal component analysis is used in order to identify and eliminate unimportant attributes within the data and the aggregation of the remaining attributes. By combining Kaiser's rule and the Broken-stick model, four principal components are selected for subsequent data modelling. Cluster analysis is used in order to identify associations and structure within the data. A hierarchy of cluster structures is constructed with two, three, four and five clusters in 21-dimensional data space. The main differences between clusters are described in this paper.

Keywords: Principal Component Analysis, Cluster Analysis, Census Data, Household Structures

e-ISSN: 2352-6343
PID article: http://hdl.handle.net/10622/23526343-2015-0004?locatt=view:master

The article can be downloaded from here.

1 INTRODUCTION

The opportunities to explore household and family patterns in new ways as a result of the emergence of new data resources providing large amounts of individual level historical microdata, sometimes covering entire countries, has been commented upon by Steven Ruggles (2012). One approach advocated by Ruggles is to undertake analyses of spatial variation, using the greater and finer geographical coverage of these new data resources to illustrate complexities and differences that single place studies cannot.[1] As one strand of a larger multi-national JISC-funded project,[2] this paper does exactly that. It explores spatial variations and patterns in household structure across mid-Victorian England and Wales in terms of socio-economic indicators, by applying multi-dimensional analysis techniques to historical geo-referenced census data. However, in so doing, it specifically does not address the decline of patriarchal family forms in Europe and beyond, a topic that Ruggles specifically suggests that these new data resources be used to address (Ruggles 2012). In part, this is because it is an analysis of just a single census year, and thus change over time cannot be detected. Moreover, this is a study of variations in *household* form rather than a study of evolving *family* systems. The two are rather different. Thus, while this research includes co-residential kinship structures as part of its analysis, it paints with a much broader brush. Moving the focus from *family* to *household* and then to *parish*, this study marshals a wide range of indices, familial and non-familial alike, in order to try and understand how the composite households and their inhabitants within one locality or place (in this case the parish) are similar or different from those in the places which surround them. In this sense, the goal is to better understand how variations at the household and parish levels contribute to broader regional differences and variations. Are households in the north, south, east or west essentially the same in mid-Victorian England and Wales, or can we detect differences at a regional level between them?

To date, there have been relatively few studies of geographical variations in historical household structure in England and Wales. Those that have been attempted have been relatively inconclusive due to a basic lack of detailed data in order to fully investigate the subject, mainly because they have had to resort to the use of aggregated census data resulting in a lack of spatial granularity and detail, or partial sources for pre-census periods (Wall 1977; Schürer 1992). Since the publication of *Household and Family in Past Time* in 1972, the common orthodoxy which has developed is that the households of the past in England and Wales were predominately nuclear in terms of family form and varied little over space (and time) (Laslett 1972; Laslett 1983).[3] This was summarised by Wall in 1983 as follows: "The basic structure of English households in the pre-industrial era is now well known. Households were small. The majority contained fewer than five persons and membership was customarily confined to parents and their unmarried children" (Wall 1983). However, despite this bold statement, any systematic attempt to consider regional variation has been mainly absent. Curiously, when Peter Laslett presented his initial findings on English historical household structure in the journal *Population Studies* in 1969 the article was entitled 'Part I'.[4] The second instalment, to be published later in the same journal, was to "describe and analyse variations in mean household size by region and by period" (Laslett 1969). But 'Part II' never appeared, it seems, primarily because there was no story to tell.

The conventional view that household structure varied little historically, has in part been re-enforced by a number of demographic studies that have emphasised the homogeneity of England's demographic experience rather than its variance – especially in comparison with other European countries (Wrigley & Schofield 1983; Wrigley 1985). Reviewing Teitelbaum's study of fertility decline in England and Wales, Laslett commented that it portrayed the demographic experience of the English like "the red coats on parade in front of Buckingham Palace, every unit in step with every other, and all changing direction at the same time" (Laslett 1985; Teitelbaum 1984; cf. Garrett, Reid, Schürer & Szreter 2001). However, we are still left with two basic problems: how much of this seemingly homogeneity is a factor of either, first, the size of the units under observation; or second, the range of variables under consideration. Teitelbaum's

[1] Ruggles (2012) proposes that studies using the newly available large data sources should use demographically appropriate measures, study spatial variation in families and households and study long-run historical changes. (p.424).

[2] The title of the JISC-funded project was "Mining Microdata: economic opportunity and spatial mobility in Britain, Canada and The United States, 1850-1911". This was undertaken jointly with the University of Alberta, University of Montreal, University of Guelph (all Canada), the Minnesota Population Center at the University of Minnesota (USA). Details are available at http://www.miningmicrodata.org/. Details on the Digging into Data research programme are at: http://www.diggingintodata.org/.

[3] The traditional picture for England and Wales varies dramatically to that recently portrayed by Szołtysek, Gruber, Klüsener & Goldstein (2014) in which they suggest a distinct north/south division with greater household complexity in the north and with disparities being explained by agriculture, fertility and differences in age structures.

[4] The sub-title, rarely cited is 'Part I. Mean Household Size in England since the Sixteenth Century'. See Laslett (1969).

study of fertility decline geographically focused on the 50 or so administrative historic county units of England and Wales. Widening the scope to 614 'registration' districts used in England and Wales in the nineteenth-century, Woods has demonstrated considerably more geographic variation in relation to mortality (Woods & Shelton 1997; Woods 2000). What would the situation look like if the telescope lens was amplified not just 10-fold, from 50 to 600 units, but over 3000-fold, to 17,000 units? And, from a household perspective, would homogeneity persist if we broadened our focus beyond size and the presence or otherwise of co-resident kin, to include summary measures on servants, lodgers, occupational concentration, isonomy, migration and so on? By significantly changing the focus of the investigation, in terms of both the geographic scope and the thematic range, this research will test the notion of homogeneity in household structure and produce a new typology of parish-based regional variation.

In order to do this, this paper will examine variations in household structure by using complete count, individual level, census data for 1881. In all, some 25 million person records have been aggregated at household and parish levels and then examined applying principal component analysis and cluster analysis. Principal component analysis (PCA) is one of the most common techniques used to describe patterns of variation in multi-dimensional data (Gorban & Zinovyev 2009). Moreover, PCA is recognised as one of the more robust ways to identify and carry out dimensionality reduction, which in turn, allows the selection of the most informative features (Abdi & Williams 2010). Cluster analysis is a tool for discovering key associations and structures within the data and typology development (MacQueen1967). Within this research, the analysis and visualisation of multi-dimensional data has been conducted using the ViDaExpert application (Zinovyev 2000). This software allows users to construct simple visual representations of the dataset in order to explore its intrinsic patterns and regularities.

The paper is structured as follows: section 2 presents a description of the data; section 3 considers the results of the PCA including the elimination of unimportant features and the aggregation of attributes, the selection of the number of principal components, the contribution of the data attributes to the principal components and data visualisation; section 4 presents the results of the cluster analysis with visualisation of two-, three-, four- and five- cluster structures within the data; section 5 presents conclusions.

2 DATA DESCRIPTION

The dataset is derived from the individual level census data from the 1881 of England and Wales (Schürer & Woollard 2000; Schürer & Woollard 2002). From this some 25 million person records were aggregated at household and then parish level. The resulting dataset used in this analysis contains 13,390 objects, essentially discrete parish-level geographical entities, each with 45 measured attributes. The set of attributes includes two basic types: the first are those providing a range of socio-economic summary measures derived from the underlying data relating to the respective parishes; the second are additional locational reference characteristics, used for data interpretation and visualisation. The dataset contains 33 main attributes and 12 additional attributes. These are listed in Table 1.

Table 1 *List of the data attributes*

Main attributes		
1	*HHsize*	Mean household size
2	*SolitaryMHH*	% of households headed by a solitary male
3	*SolitaryFHH*	% of households headed by a solitary female
4	*HH_with_kin*	% of households with residential kin
5	*HH_with_servt*	% of households with residential servants
6	*HH_with_inmates*	% of households with non-family members
7	*WorkingF25+*	% of females aged 25+ who are working
8	*Working20+*	% aged 20+ who are working
9	*Working<=14*	% aged 14 and less who are working

Main attributes		
10	*Working55+*	% of males aged 55+ who are working
11	*Males_in_agric*	% of males aged 25+ working in agriculture
12	*Native*	% who are native (born in same county)
13	*Foreign*	% who are born overseas
14	*Scottish*	% born in Scotland
15	*Irish*	% born in Ireland
16	*HHsize6+*	% of households with 6 or more offspring
17	*No_par<=5*	% aged 5 or less living without parents
18	*Sing_Par<=5*	% aged 5 or less living with a single parent
19	*Live_with_par15-16*	% aged 15-16 living in the parental home
20	*Live_with_par17-18*	% aged 17-18 living in the parental home
21	*Live_with_par19-20*	% aged 19-20 living in the parental home
22	*Live_with_par21-22*	% aged 21-22 living in the parental home
23	*With_older_sibs*	% aged 25+ living with siblings aged 25+
24	*Aunt/uncle*	% living with aunts or uncles
25	*Nephew/niece*	% living with nieces or nephews
26	*Cousins*	% living with cousins
27	*Grandparents*	% living with grandparents
28	*Grandchildren*	% living with grandchildren
29	*Occ_similarity*	Measure of occupation concentration
30	*Name_similarity*	Measure of surname heterogeneity
31	*Blind*	% blind
32	*Deaf*	% deaf
33	*Mental*	% with mental disability
Additional attributes		
34	*Standardparish*	Name of place
35	*Country*	Country
36	*Division*	Census Division
37	*RC*	Census Registration County
38	*RC_ref*	Census Registration County ref
39	*RD*	Census Registration District
40	*RD_ref*	Census Registration District ref
41	*Area*	Area of parish unit
42	*Aggpop*	Population size of parish unit
43	*Density*	Population density of parish unit
44	*X_centroid*	X coordinate of parish unit
45	*Y_centroid*	Y coordinate of parish unit

3 DATA AGGREGATION

The data aggregation process includes both the aggregation of attributes and the elimination of unimportant features. The aggregation of attributes is based on PCA techniques and correlation analysis. To identify attributes with similarities, the contribution of the data attributes to the four principal components was analysed. This suggested that there are six groups of attributes with equal signatures:

1 – *SolitaryMHH* and *SolitaryFHH*
2 – *WorkingF25+* and *Working20+*
3 – *Live_with_par17-18*; *Live_with_par19-20* and *Live_with_par21-22*
4 – *Aunt/uncle, Nephew/niece* and *Cousins*
5 – *Grandchildren* and *Grandparents*
6 – *Occ_similarity* and *Name_similarity*.

The results demonstrate a strong correlation between attributes. Taking into account the contribution of the data attributes to the principal components and correlation coefficients between attributes, it was possible to create the following aggregate attributes:

1 – *SolitaryHH*
2 – *Working20+*
3 – *Live_with_par17-22*
4 – *Distant_relatives*
5 – *Gdchildren/Gdparents*
6 – *Occ/names_similarity*

As a result of data aggregation, the number of attributes was reduced to 25 (from 33).

The elimination of unimportant features is based upon a PCA definition of unimportant attributes. The criterion for the definition of unimportant attributes is Kaiser's rule for eigenvector of the principal components: $z_i^2 < \frac{1}{n}\sum_{i=1}^{n} z_i^2$, where $i = \overline{1,n}$; n – is a number of attributes; λ_i – is a value of i-th attribute in eigenvector. The attributes that have values less than the average value for all principal components are excluded. The analysis of the principal components showed that there were four attributes which could be deleted from further analysis: *P5singpar, Blind, Deaf* and *Mental*. Consequently, after elimination of unimportant features, the dataset contained only 21 attributes.

4 PRINCIPAL COMPONENT ANALYSIS

PCA is one of the most common techniques used to describe patterns of variation within a multi-dimensional dataset, and is one of the simplest and robust ways of doing dimensionality reduction. PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Peres-Neto, Jackson & Somers 2005). The number of principal components is always less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance and each subsequent component, respectively, has the highest variance possible under the constraint that it be orthogonal to the preceding components.

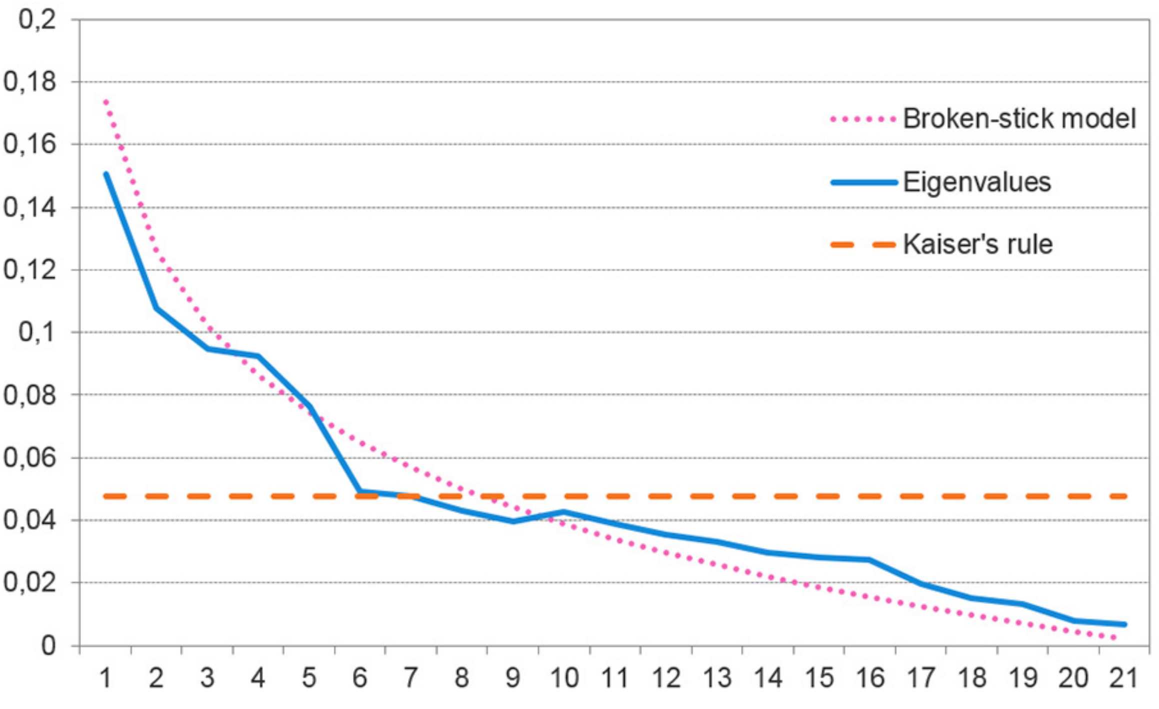
4.1 SELECTION OF THE NUMBER OF PRINCIPAL COMPONENTS

One of the greatest challenges in providing a meaningful interpretation of multi-dimensional data using PCA is determining the number of principal components. There are various methods and stopping rules used to identify the number of principal components. In selecting the number of principal components we applied the most commonly used method, namely Kaiser's rule and the Broken-stick model based on eigenvalues of components. According to Kaiser's rule, the components that have eigen values greater than the average value are retained for interpretation: $\lambda_i > \frac{1}{n}\sum_{i=1}^{n}\lambda_i$, where $i = \overline{1,n}$; n – is a number of components; λ_i – is a eigenvalue of i -th component. The concept underlying the Broken-stick model is that if a stick is randomly broken into n pieces, l_1 would be the average size of the largest piece in each set of broken sticks; l_2 would be the average size of the second largest piece, and so on. The number n equals the number of components and the total amount of variation across all components. The proportion of total variance associated with the eigenvalue for i -th component under the broken-stick model is obtained by: $l_i = \frac{1}{n}\sum_{j=i}^{n}\frac{1}{j}$. If the i -th component has an eigenvalue larger than l_i , then the component is retained. Initially, four principal components were identified.

Principal components for a reduced number of data attributes were selected based on combination of Kaiser's rule and the Broken-stick model. Figure 1 illustrates the eigenvalues of components.

As can be seen from Figure 1, Kaiser's rule determines five principal components – eigenvalues of first five components are significantly greater than the average value. The Broken-stick model gives three principal components – the line of Broken-stick model cuts the eigenvalues of first three components. In addition, the spectral gap (i.e. the distance between eigenvalues) separates the first component significantly, and the second, third, fourth and fifth components from other components. Consequently, for reduced data attributes four principal components were identified: PC1, PC2, PC3 and PC4.

Figure 1 *Eigenvalues of components for reduced data attributes*



4.2 CONTRIBUTION OF THE DATA ATTRIBUTES TO THE PRINCIPAL COMPONENTS

The contribution of the reduced data attributes to principal components is represented in Figures 2-5.

The first principal component (PC1, Figure 2) is characterised by the following attributes: moderately large *household size*; high proportions of both *households with residential kin* and *households with residential servants*; a high percentage of *males working in agriculture*; a strong negative correlation with the percentage of *households with six or more offspring* and also *children (ages from 15 to 22) living in the parental home*; a high proportion of *children aged 5 or less living without parents*; high proportions *living with siblings, aunts or uncles, nieces or nephews, cousins, grandparents* and *grandchildren*; and high levels of *occupation concentration* and *surname concentration* (i.e. relatively low surname heterogeneity). In combination, these components suggest rural parishes dominated by a single source of employment (agriculture) with large families, but where residential (extended) kin and servants are an important element of overall household size pro rata to offspring. Strong surname concentration may also indicate a less mobile population.

The second principal component (PC2, Figure 3) is characterised by the following attributes: relatively small *household size*; a high percentage of *households with residential servants* and *households with non-family members*; low proportions of *males working in agriculture*; relatively low proportions *native born* and high percentages *born overseas, born in Scotland* and *born in Ireland*; low proportions of households with *six or more offspring*; high proportions of *children aged 5 or less living without parents*; low proportions of *children (ages from 15 to 22) living in the parental home*; high proportions of households with members *living with siblings, aunts or uncles, nieces or nephews* and *cousins*; together with high levels of *occupation concentration* and *surname concentration*. In combination, these components suggest mainly inner urban parishes with a mobile population and varied economy/occupation structure, with relatively small households, but where residential (extended) kin, boarders. lodgers and servants are an important element of overall household size pro rata to offspring.

The third principal component (PC3, Figure 4) is characterised by the following attributes: moderately large *household size*; high proportion of *households with residential kin*; low proportions of *households with residential servants*; low percentage of *males working in agriculture*; high proportion of households with *six or more offspring;* high percentages of *children (ages from 15 to 22) living in the parental home*; high percentages l*iving with siblings, aunts or uncles, nieces or nephews, cousins, grandparents* and *grandchildren*; and low levels of *occupation concentration* and *surname concentration*. In combination, these components suggest parishes with a fairly mixed economy/occupational structure, yet which are not urban areas with a high migrant component – maybe smaller market towns – with large families where both residential kin and the retention of children in the household are important, yet servants less so.

The fourth principal component (PC4, Figure 5) is characterised by the following attributes: *large household size*; low proportions of *households with residential kin*; high proportions of *households with residential servants*; low percentages of *males working in agriculture*; high proportions of *households with six or more offspring*; low proportions of *children (ages from 17 to 22) living in the parental home;* low proportions *living with siblings, aunts or uncles, nieces or nephews, cousins, grandparents* and *grandchildren*; and relatively high levels of *occupation concentration* and *surname concentration*. In combination, these components suggest non-agricultural parishes yet with relatively little variation in the local econo my/occupational structure and a fairly 'stable' non-migratory population, with large households in which young children are a key element (suggesting maybe higher fertility). These characteristics could indicate mining and similar 'mono-culture' communities.

Figure 2 *Contribution of the reduced data attributes to PC1*

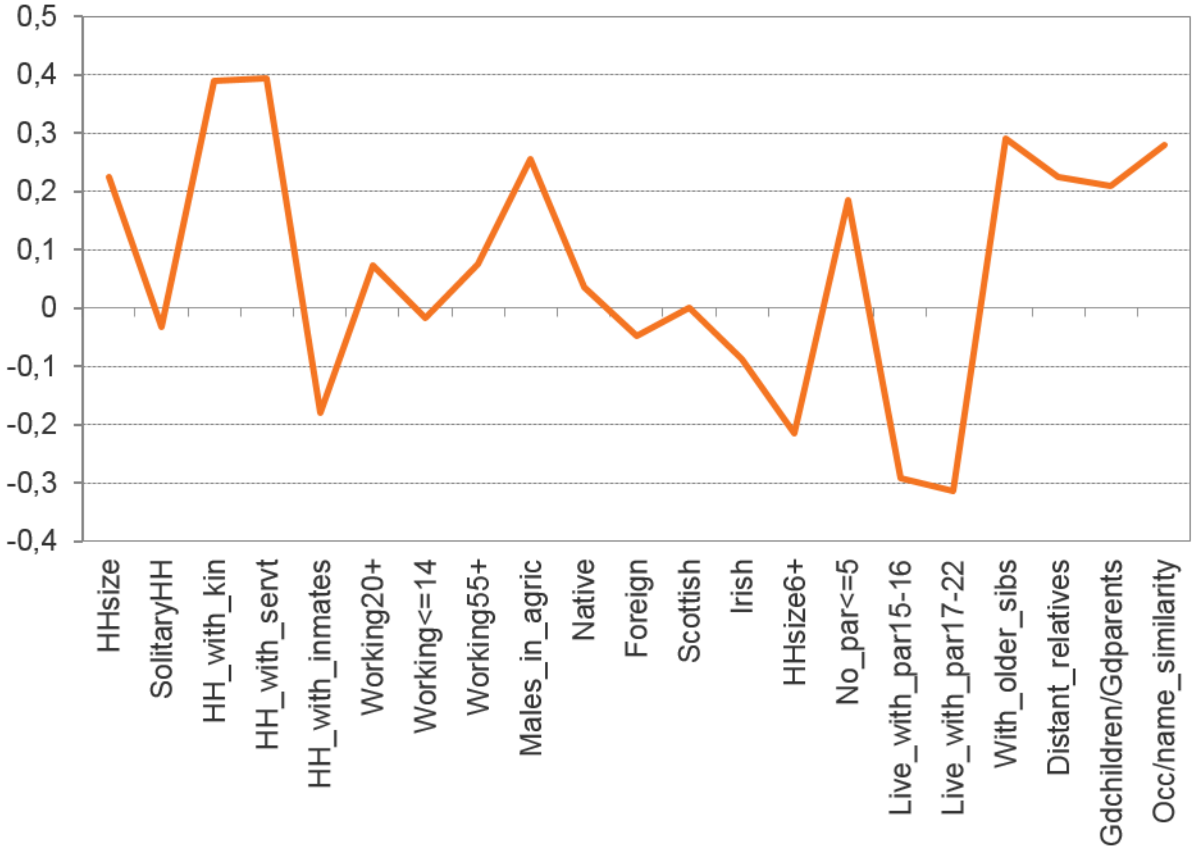


Figure 3 *Contribution of the reduced data attributes to PC2*

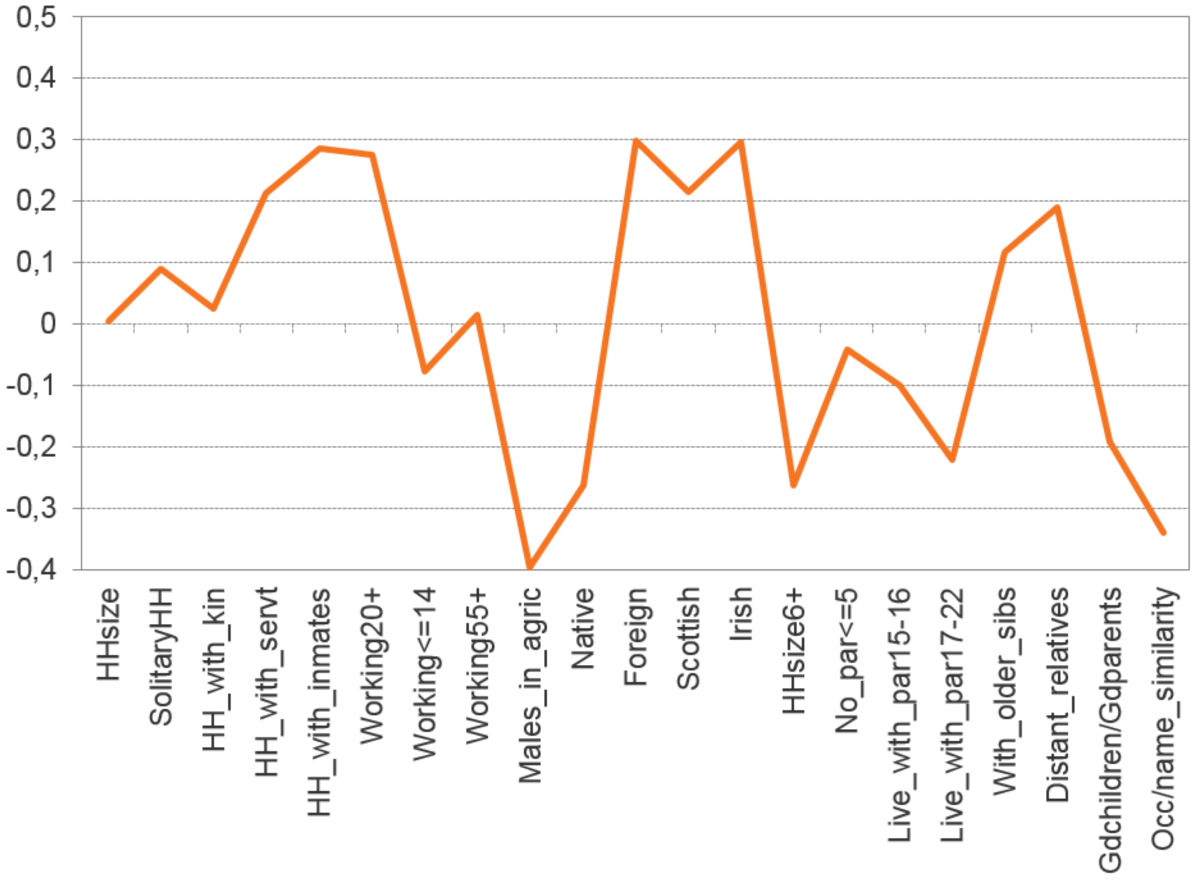


Figure 4 *Contribution of the reduced data attributes to PC3*

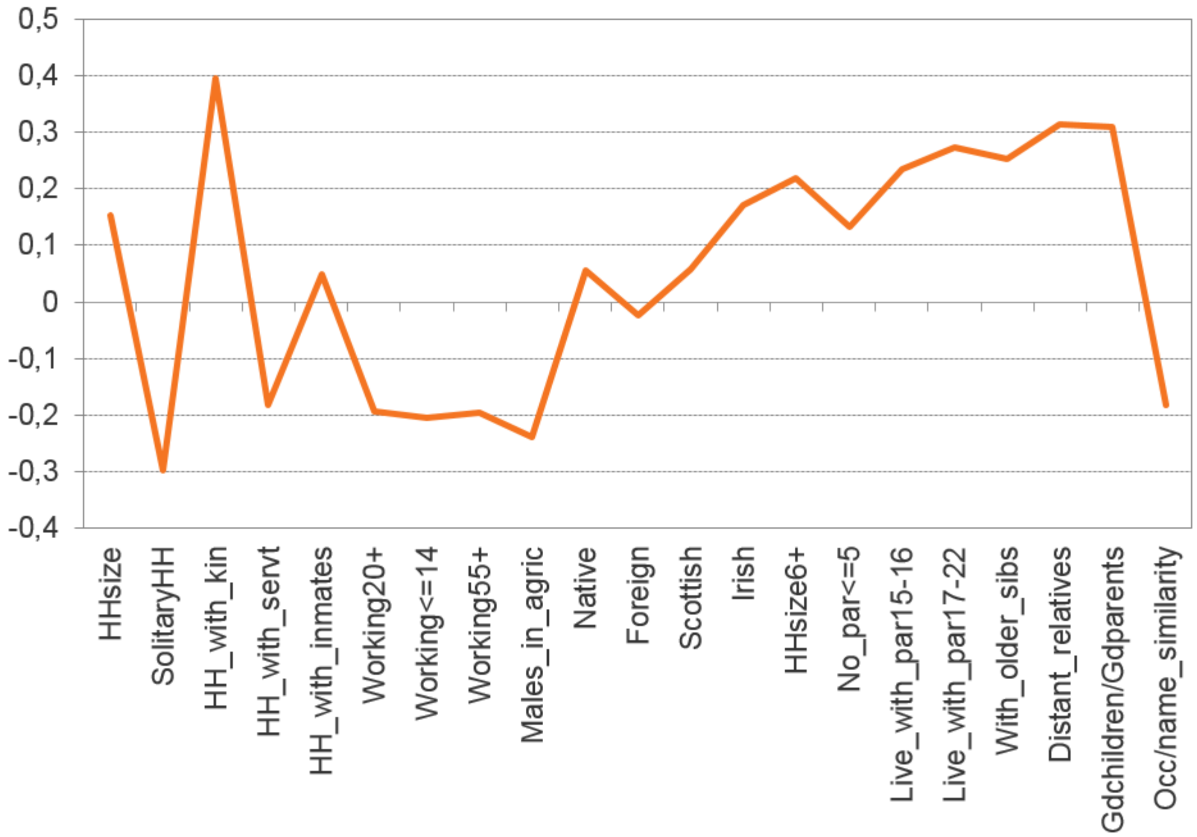
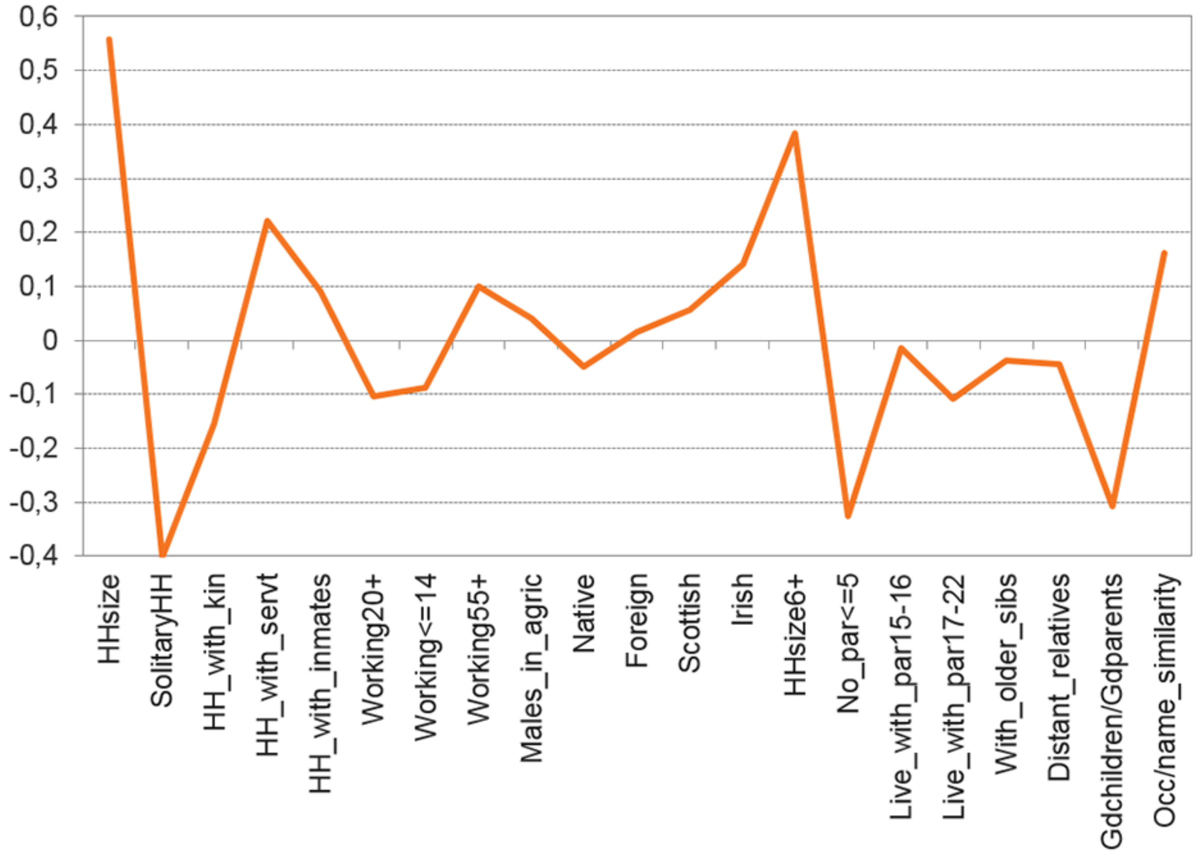


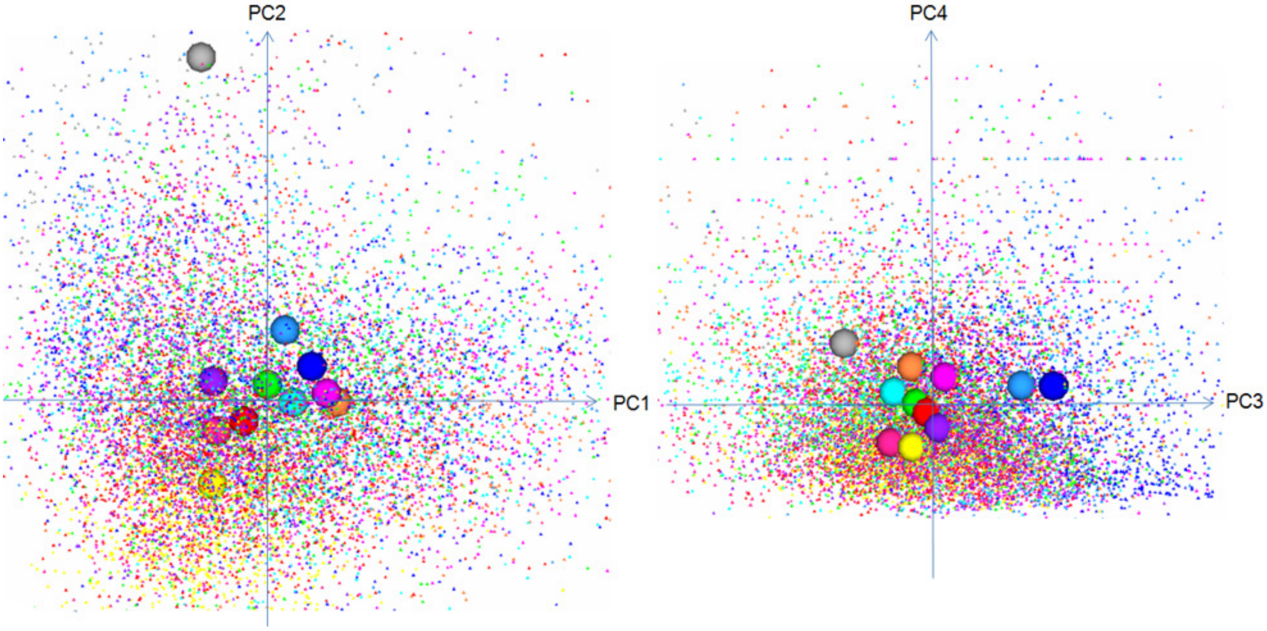
Figure 5 *Contribution of the reduced data attributes to PC4*



4.3 DATA DISTRIBUTION ON THE PRINCIPAL COMPONENTS

The data can be divided into eleven groups according to where the objects (parishes) are located in terms of Standard Regions. These are: group 1 (blue) – North, 941 objects; group 2 (rose) – Yorkshire, 1407 objects; group 3 (light blue) – North Western, 874 objects; group 4 (turquoise) – North Midland, 1546 objects; group 5 (brown) – Monmouth/Wales, 1093 objects; group 6 (green) – West Midland, 1515 objects; group 7 (red) – South Western, 1696 objects; group 8 (crimson) – South Midland, 1318 objects; group 9 (purple) – South East, 1371 objects; group 10 (yellow) – Eastern, 1473 objects; group 11 (grey) – London, 156 objects. Figure 6 shows the visualisation of these eleven standard geographic regions on the PCA plot.

Figure 6 *Visualisation of geographic regions on the PCA plot*



As can be seen from Figure 6, regions such as Wales, Yorkshire, North, North Midland, West Midland, South Western, South Midland and South East are mainly distributed along the first principal component (PC1), while Eastern and London are associated with the second principal component (PC2). The North, North Western and London differ from other regions by the third principal component (PC3), while Wales, Yorkshire, North Midland, West Midland, South Western, South Midland, South East and Eastern are distributed along the fourth principal component (PC4).

According to values of principal component projections, the data were divided into five groups. The results of data grouping are represented in Table 2.

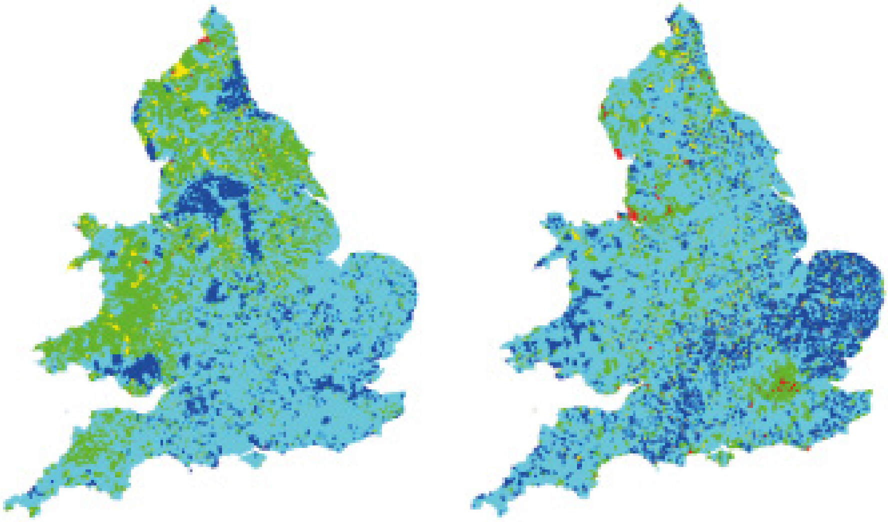
Table 2 *Data grouping according to values of principal component projections*

GROUPS	PC1	PC2	PC3	PC4
Group 1 (blue) n. of objects	-4.482, -1.803 8,185	-5.948, -1.004 3,168	-11.681,-3.888 106	-8.334, -2.927 153
Group 2 (light blue) n. of objects	-1.803, 0.874 1,679	-1.004, 1.467 8,417	-3.888, -1.342 1,848	-2.927, -0.232 6,029
Group 3 (green) n. of objects	0.874, 3.551 3,014	1.467, 3.943 1,541	-1.342, 1.243 9,208	-0.232, 2.474 6,567
Group 4 (yellow) n. of objects	0.551, 6.245 440	3.943, 6.426 225	1.243, 3.828 2,143	2.474, 5.217 586
Group 5 (red) n. of objects	6.245, 14.261 72	6.426, 16.298 39	3.828, 8.997 85	5.217, 13.271 55

Figure 7 displays the visualisation of the projections on the first and second principal components based on geographic coordinates. As can be seen from the visualisation of the first component (Figure 7, left), the low values of projections (light blue points) are dominant in the southern part of England; the high values of projections (green, yellow and red points) dominate in the northern part of England and in Wales. Besides, the lowest values (blue points) are concentrated as big cities across the country.

The visualisation of the second component (Figure 7, right) demonstrates that the low values of projections (blue points) also are dominant in the southern part of England; the high values of projections (light blue points) dominate in northern part of England and in Wales. However, the highest values (yellow and red points) are concentrated in larger towns and cities across the country in the southern part of England; the high values of projections (green, yellow and red points) dominate in the northern part of England and in Wales. Besides, the lowest values (blue points) are concentrated as big cities across the country.

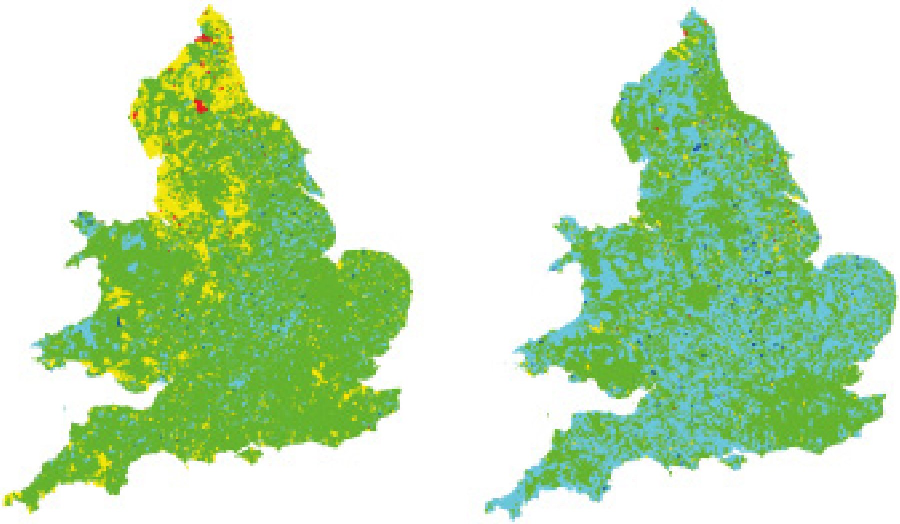
Figure 7 *Visualisation of the projections on the first and second principal components on the geographic coordinates*



Note: *see Table 2 and text for the colour assignment*

Figure 8 displays the visualisation of the projections on the third and fourth principal components on the geographic coordinates. The visualisation of the third component (Figure 8, left) illustrates that objects with high values of projections (yellow and red points) are observed primarily in the north of the country, while objects with low values of projections are occur mainly in central England (light blue and blue points) and the south (green points). Also, we can notice that objects with high values are concentrated in large towns and cities. The visualisation of the fourth component (Figure 8, right) illustrates that objects with high values of projections (yellow and red points) occur mainly in the north of the country, while objects with low values of projections are observed in the central and southern regions. We also can see objects with higher values in the larger towns and cities across the country.

Figure 8 *Visualisation of the projections on the third and fourth principal components on the geographic coordinates*



Note: see Table 2 and text for the colour assignment

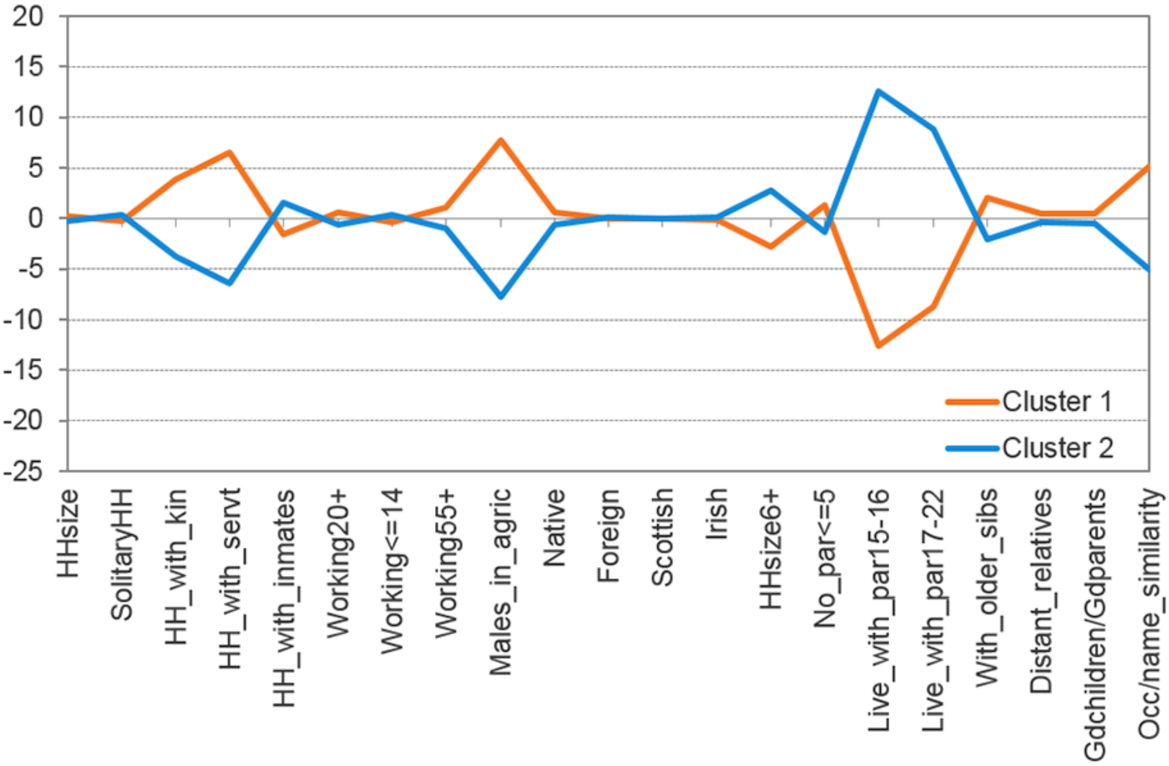
5 CLUSTER ANALYSIS

Cluster analysis is a tool for discovering and identifying associations and structure within the data and typology development (MacQueen 1967). Cluster analysis provides insight into the data by dividing the dataset of objects into groups (clusters) of objects, such that objects in a cluster are more similar to each other than to objects in other clusters. At present, there are many various clustering algorithms which are categorized based on their cluster model (Jain & Dubes 1988). In this research, for cluster analysis of census data the centroid-based clustering method is used. *K*-means is a well-known and widely used clustering method which aims to partition objects based on attributes into *k* clusters. The *k*-means clustering is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The centroid can be interpreted as a prototypical point for this cluster. The *K*-means method has two key features: 1) Euclidean distance is used as a metric and variance is used as a measure of cluster scatter; 2) the number of clusters (*k*) is an input parameter which should be specified in advance. For the *k*-means clustering method the most important and difficult question is the identification of the number of clusters that should be considered. In this case, in order to determine the number of clusters the PCA technique was used: the number of clusters being dependent upon the number of principal components. Thus, referring back to the previous discussion, the first component forms two clusters, second component forms three clusters, and so on. According to the eigenvalues of components (Figure 1 above) there are 1-4 principal components. This means that the data has 2-5-cluster structures, where *k*=5, is the maximum number of informative (significant) clusters.

5.1 TWO-CLUSTER STRUCTURE

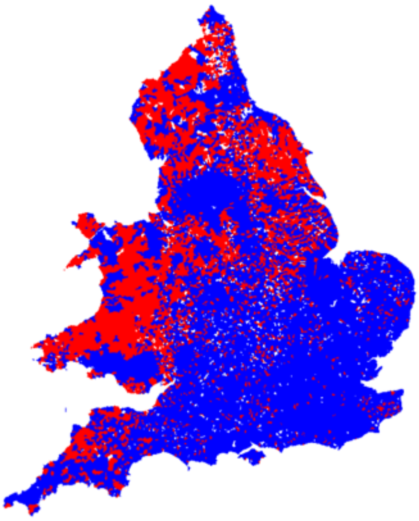
In the two-cluster structure (*k*=2) cluster 1 (blue) has 9,118 objects and cluster 2 (orange) has 4,272 objects.The difference between clusters is identified by the standard deviation of cluster averages of attributes. Figure 9 shows the distribution of the clustered data on the attributes in two-cluster structure.

Figure 9 *Distribution of the clustered data on the attributes in two-cluster structure*



As can be seen, the two clusters differ significantly on such characteristics as: *households with residential kin*; *households with residential servants*; *males working in agriculture; households with six or more offspring*; *children living in the parental home* and *occupation/surname concentration*. Cluster 1 is characterized by *high proportions of households with residential kin*, *households with residential servants* and *males working in agriculture*; low proportions of *children living in the parental home*; slightly lower values of *households with six or more offspring* and moderate proportions of *occupation/surname concentration*. Cluster 2 is the mirror image of this pattern. The distribution of the clustered data on the regions in two-cluster structure is represented in Figure 10. As can be seen, the elements of Cluster 1 dominate southern England, and run through the midlands, while the elements of cluster 2 dominate in east and north Yorkshire, the north-west around Cumbria, south Lancashire, Wales, and curiously, Devon.

Figure 10 *Two-cluster structure on the geographic coordinates*

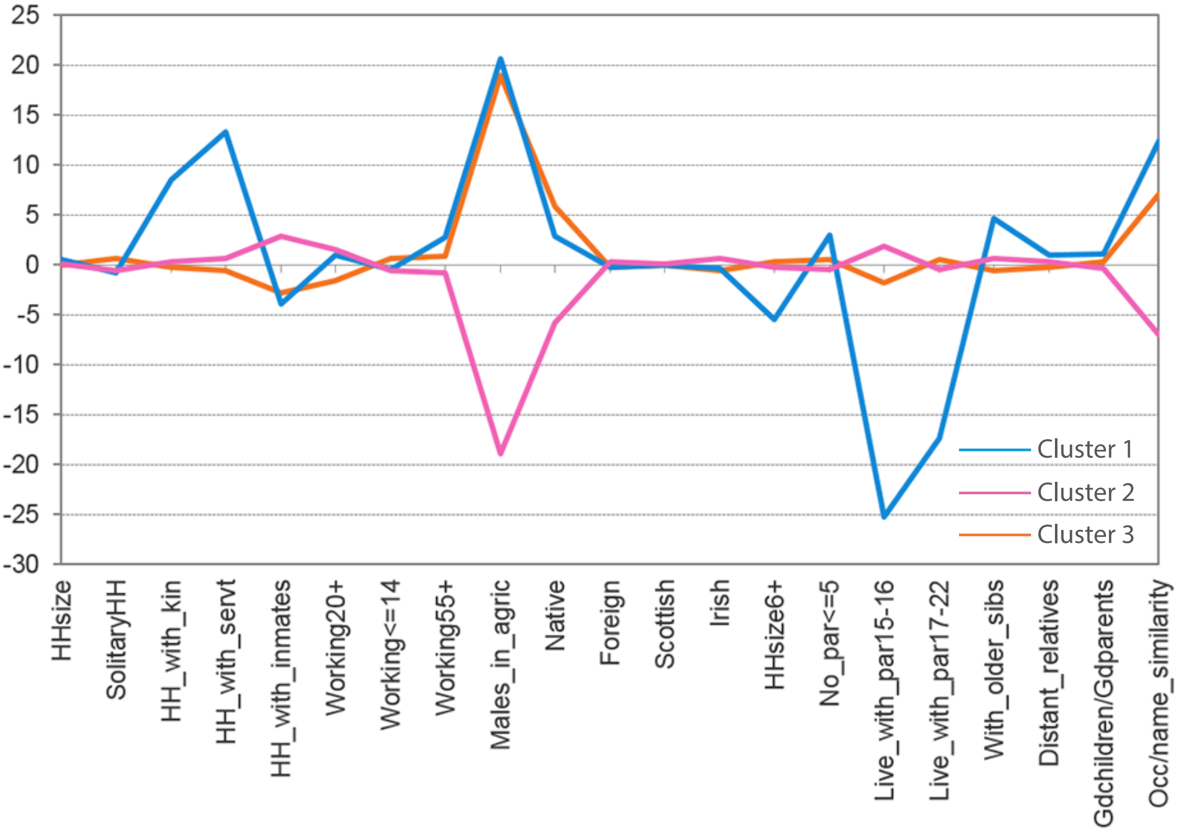


Note: Cluster 1 = blue, Cluster 2=red

5.2 THREE-CLUSTER STRUCTURE

The clusters within the three-cluster structure (*k*=3) are: cluster 1 (blue) with 6,662 objects, cluster 2 (pink) with 3,353 objects and cluster 3 (orange) with 3,375 objects. Figure 11 shows the distribution of the clustered data on the attributes in three-cluster structure.

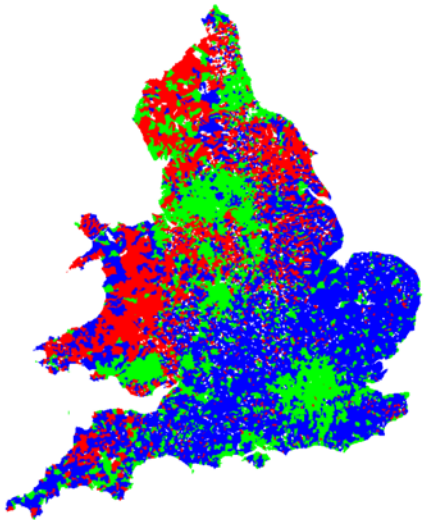
Figure 11 *Distribution of the clustered data on the attributes in three-cluster structure*



As can be seen, the three clusters are significantly different on such characteristics as: *households with residential kin*; *households with residential servants*; *males working in agriculture*; *children living in the parental home*. Each cluster is different and cluster 3 is dramatically different from the other two clusters. Cluster 3 is characterized by high proportions of *households with residential kin*, *households with residential servants* and *males working in agriculture*; and low proportions of *children living in the parental home*; *households with six or more offspring*; as well as a high value for *occupation concentration* and *surname similarity*.

In contrast, clusters 1 and 2 tend to differ from cluster 3 on all the key attributes mentioned above, with the exception of cluster 1 having similar experience in *males working in agriculture.* In contrast, cluster 2 stands out as having low levels of *males working in agriculture* and a correspondingly low value for *occupation concentration* and *surname similarity*, which in combination would suggest that this cluster is mainly urban. Figure 12 illustrates the geographical distribution of the separate clusters. As can be seen, in combination these nuance the two cluster model described earlier. Cluster 2 in the three cluster structure essentially removes the predominantly urban places from cluster 1 of the two cluster structure discussed earlier, leaving a basic north-south divide represented by clusters 1 and 3 – roughly diagonal Severn-Wash line – although north Devon again stands out.

Figure 12 *Three-cluster structure on the geographic coordinates*

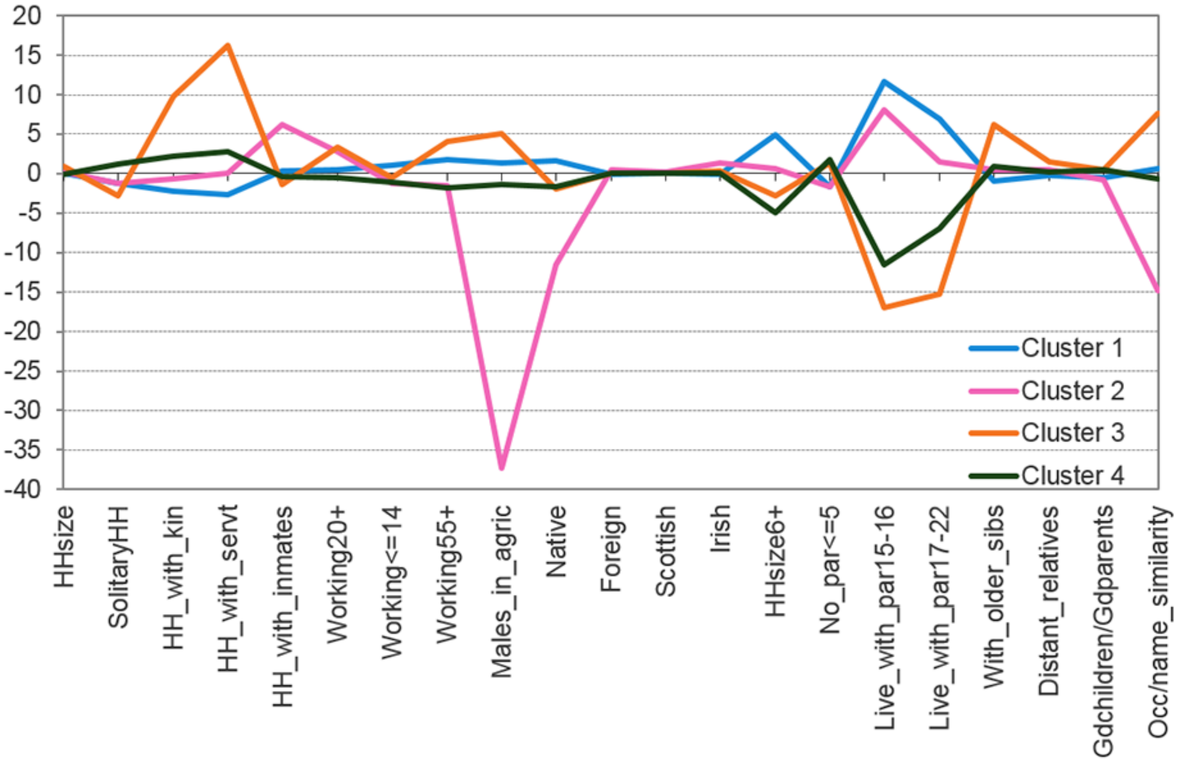


Note: *Cluster 1 = blue, Cluster 2 = green, Cluster 3 = red*

5.3 FOUR-CLUSTER STRUCTURE

The four-cluster structure (*k*=4) is as follows: cluster 1 (blue) with 4,350 objects, cluster 2 (pink) with 2,992 objects, cluster 3 (green) with 4,096 objects, and cluster 4 (orange) with 1,952 objects. Figure 13 shows the distribution of the clustered data in relation to the attributes within the four-cluster structure.

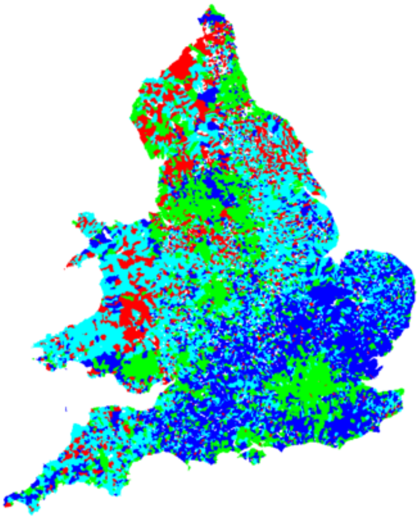
Figure 13 *Distribution of the clustered data on the attributes in four-cluster structure*



As can be seen, the four clusters differ considerably around the following key characteristics: *households with residential kin*; *households with residential servants*; *households with unrelated persons*; *males working in agriculture*; *households with 6 or more offspring*; *children living in the parental home* and *occupation concentration* and *surname similarity*. Cluster 1 is characterised by high proportions of *households with six or more offspring* and high retention of children within the parental home, and conversely, low proportions of *households with servants*. Cluster 2 shows what might be seen as common characteristics of urban populations: significantly low proportions of *males working in agriculture* together with a low value for *natives* and very low value for *occupation concentration* and *surname similarity*. Also this cluster displays relatively high proportions of *households with unrelated persons* (boarders and lodgers) and *children living in the parental home* (aged 15-16). Cluster 3 is in some respects the mirror image of Cluster 1. It has low proportions of *households with six or more offspring*, relatively low retention of *children living in the parental home*, together with relatively high proportions of *households with residential kin* and *servants*. Lastly, cluster 4 is conversely characterised by high proportions of *households with residential kin* and *servants*, together with relatively high proportions of *males working in agriculture* and *elderly workers*. Equally, the proportions of *households with six or more offspring* and *children living in the parental home* is low, while the proportion *households with elderly siblings* living together is relatively higher and the value for *occupation concentration* and *surname similarity* is comparatively very high. These characteristics suggest rural places dominated by mono-cultures.

Figure 14 maps the geographic distribution of the 4 clusters. This illustrates, as already indicated, that cluster 2 within the four-cluster structure is primarily composed of larger urban communities, distributed across the country. In contrast, cluster 1 features mainly in southern rural England, but interestingly, moving from the three to four cluster structure suggests a split between the south-west (Cornwall and Devon), and the rest of southern England (south of the Severn-Wash) line. The south-west joins cluster 3 in this model, in a mainly rural northern England/Wale grouping, but within which parts of East Anglia are also represented. Lastly cluster 4 parishes are located mainly in the north of England, with especially heavy concentrations in south Lancashire, Northumberland, Durham and East Yorkshire. In part, it is tempting to suggest that this cluster could be influenced by the existence of mining industries, but Figure 14 indicates that this is not exclusively mining.

Figure 14 *Four-cluster structure on the geographic coordinates*

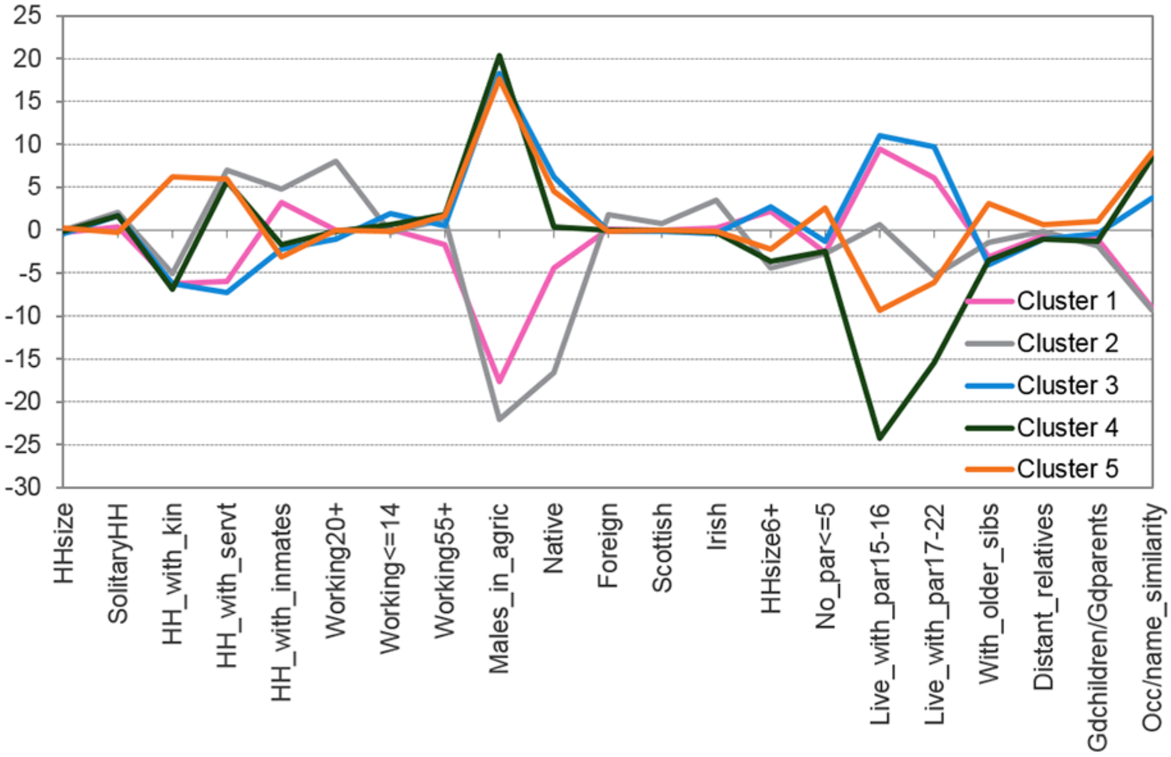


Note: Cluster 1 = dark blue, Cluster 2 = green, Cluster 3 = light blue, Cluster 4 = red

5.4 FIVE-CLUSTER STRUCTURE

The breakdown of the five-cluster structure (*k*=5) is as follows: cluster 1 (blue) with 4,789 objects, cluster 2 (pink) with 3,462 objects, cluster 3 (gray) with 543 objects, cluster 4 (green) with 2,511, and cluster 5 (orange) with 2,085 objects. Figure 15 shows the distribution of the clustered data in relation to the attributes in five-cluster structure. As can be seen, the five clusters different significantly around the following characteristics: *households with residential kin*; *households with residential servants*; *households with unrelated persons*; *population ages 20 are working*; *males working in agriculture*; *native*; *households with 6 or more offspring*; *children living in the parental home* and *occupation concentration* and *surname similarity*.

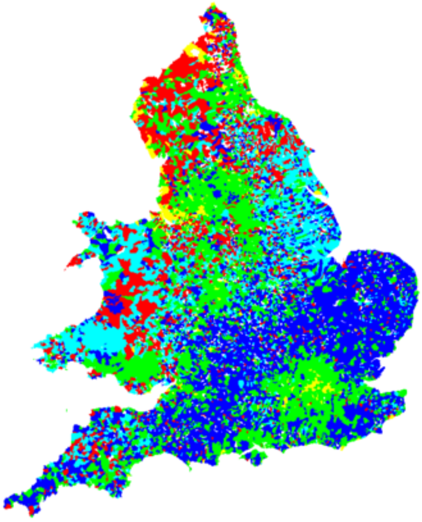
Figure 15 *Distribution of the clustered data on the attributes in five-cluster structure*



Cluster 1 is characterised by moderately low proportions of *households with residential kin* and *servants*; low proportions of *households with unrelated persons*; high proportions of *males working in agriculture* and *children living in the parental home*; moderately high proportions for *occupation concentration* and *surname similarity*; and slightly higher proportions of *households with 6 or more offspring.* In contrast, cluster 2 is characterized by very low proportions of both *males working in agriculture* and *natives*; moderately low proportions of *households with residential kin*; *servants* and for *occupation concentration* and *surname similarity*, together with high proportions of *children (15-22) living in the parental home*. Cluster 5 is virtually the mirror of the cluster 2 experience. Cluster 3, like cluster 2, has very low proportions of both *males working in agriculture* and *natives*, more so than cluster 2 especially in relation to *natives*; moderately low proportions of *households with residential kin* and for *occupation concentration* and *surname similarity;* slightly low proportion of *children (17-22) living in the parental home* and *households with 6 or more offspring*; yet moderately high proportions *households with residential servants*; *unrelated persons* and *those aged 20+ working*. Lastly, cluster 4, likes clusters 1 and 5, has high proportions of *males working in agriculture*; moderately high proportions of *households with residential servants* and *occupation concentration* and surname similarity; together with very low proportions of children (15-16) *living in the parental home* and to a lesser extent aged 17-22; and moderately low proportions of *households with residential kin.*

The five cluster map (Figure 16) still has cluster 1 (blue) dominating in the south of England, in a north-south divide running from the Severn to the Wash, except for the extreme south-west. In comparison to the four cluster model, the extra-metropolitan area around London falling into cluster 2 (green) is even more pronounced, especially around Surrey and Middlesex, while parts of inner London fall with cluster 3 (yellow) characterised mainly by large urban city centres, yet not exclusively so. In addition to extra-metropolitan London, cluster 4 also links to the northern counties or Durham, south Lancashire, west Yorkshire, Cheshire and down to Derby and parts of the West Midlands, as well as, Glamorgan in south Wales. This cluster would appear to represent mixed, mainly urbanised industrial economies. This is partly shadowed by cluster 5 (red) which is less urban, less industrial but is mainly northern, predominating in Cumbria, Northumbria, north Yorkshire and north Lancashire, yet with few clear concentrations. Lastly, cluster 4 (light blue) would also appear to be predominantly rural, being focused in Wales, the east of England north of the Wash, especially Lincolnshire and east Yorkshire, as well as the south-west.

Figure 16 *Five-cluster structure on the geographic coordinates*



Note: Cluster1 = dark blue, Cluster 2 = green, Cluster 3 = yellow, Cluster 4 = light blue, Cluster 5 = red

6 CONCLUSION

So what do all of these statistics and these maps tell us? Turning first to Wall's analysis of 1851 census data, aggregated by standard regions, which focused primarily on household complexity in terms of kinship, this revealed few clear patterns. However, in general terms Wales and south-west of England had the lowest levels of complexity, northern England the highest, and with eastern England being roughly in the middle (Wall 1977). Again focusing on household structural complexity, by 1981 this changed significantly, reversing in some instances. A basic dividing line could be seen running east from the Bristol Channel to the Cotswolds then turning northwards along the spine of the Pennies before heading east towards the Irish Sea below Cumbria (Wall 1982). West of this line household complexity was generally higher than to the east of the line: East Anglia recorded the lowest levels of complexity, but breaking away from this general dichotomy, London was associated with high levels of household complexity. A regional analysis of Marriage Duty Act data for the late seventeenth century, which has only patchy national coverage, revealed little in terms of clear regional variations, yet did demonstrate the distinctiveness of London and the importance of rural/urban of a potential dichotomy (Schürer 1992). Moving, to more recent trends, analysis of the 1991 census data suggests that the percentage of one person households was generally low across the western and central counties of England, slightly higher north of a line running from the rivers Mersey to Humber, including Wales, with higher percentages also recorded for London, east Sussex, Devon and Dorset. Likewise, the proportions of lone parent households were lower in the eastern counties and higher for a belt running down the centre of England, from Lancashire to Kent, as well as being high in south Wales (Champion, Wong, Rooke, Dorling, Coombes & Brunsdon 1996). More recently, Dorling and Thomas, mapping the 2001 census data for the UK suggest a growing trend towards what they term 'London and the Archipelago'. They argue that the UK is becoming more and more divided, with an imaginary line running from the rivers Severn to Humber separating a growing London metropolis from the rest of the UK. Within the London core, population is more densely concentrated, increasingly becoming younger. To the north of the line within the archipelago, are numerous centres each with their own outer areas and remoter edges. Essentially, the archipelago is an amalgam of places which have most in common in not being in the London metropolis - where, in general, population is less concentrated, often decreasing in numbers, becoming older and focusing on industries that have died or are dying (Dorling & Thomas 2004).

The analysis presented here both confirms elements of the previous work outlined above, yet adds also a much greater level of clarity. It shows that the pattern of regional variation in household structure varies in detail as different levels of complexity are considered. In part, this is like viewing a landscape through the lens of a telescope whilst gradually focusing. At a basic level, the geography of household structure is defined by a two-fold division, with a noticeable north-south divide running diagonally across the country, from the river Severn to the Wash, but taking in parts of the south midlands as well. This is not characterised by a simple urban/rural divide, as both sides of the line contain each of these elements. However, as one focuses further, urbanisation (and industrialisation) does become more of a defining feature. London, and as one focuses further, its extra-metropolitan surroundings, becomes a distinct 'region' – illustrating that the process described by Dorling and Thomas has long historical roots. North of the Severn-Wash divide, rural areas begin to segregate, with the more northerly rural areas showing distinct differences from those in the east and in Wales, with residential kin, in particular, being a key difference between these two rural types, as is the retention of children within the parental home. The evidence of this research suggests that regional variations in the patterns of residential kinship, children at home, the keeping of servants and addition of unrelated household members, such as boarders and lodgers, did exist in nineteenth-century England and Wales independent of urban and industrial drivers.

ACKNOWLEDGEMENTS

The research reported in this paper formed part of the Integrated Census Microdata (I-CeM) project, based at the Department of History, University of Essex. I-CeM was funded by the (UK) Economic and Social Research Council (award RES-061-23-1629) and supported by BrightSolid, who kindly supplied the underlying raw data. Additional work on the data was supported as part of the (Joint Information Systems Committee) JISC-funded project "Mining Microdata: economic opportunity and spatial mobility in Britain, Canada and The United States, 1850-1911". This was undertaken jointly with the University of Alberta, University of Montreal, University of Guelph (all Canada), the Minnesota Population Center at the University of Minnesota (USA). Details are available at http://www.miningmicrodata.org/. The authors would also like to thank Professor Alexander Gorban of the Department of Mathematics, University of Leicester and Andrey Zinoviev of the Institut Curie, Paris.

REFERENCES

Abdi, H. & Williams, L. (2010). Principal Components Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics,* 2(4), (pp 439-459).
DOI: 10.1002/wics.101

Champion, T., Wong, C., Rooke, A., Dorling, D., Coombes, M. & Brunsdon, C. (1996). *The Population of Britain in the 1990s. A social and economic atlas.* Oxford: Clarendon Press.

Dorling, D. & Thomas, B. (2004). *People and places. A 2001 Census atlas of the UK.* Bristol: Policy Press

Garrett, E., Reid, A., Schürer, K. & Szreter, S. (2001). *Changing Family Size in England and Wales. Place, Class and Demography, 1891-1911.* Cambridge: Cambridge University Press.

Gorban, A. N. & Zinovyev, A. Y. (2009). Principal Graphs and Manifolds, In: E.S. Olivas, J.D.M. Guererro, M.M. Sober, J.R.M. Benedito & A.J.S. Lopes (Eds.) *Handbook of Research on Machine Learning Applicat ions and Trends: Algorithms, Methods and Techniques*, (pp 28-59) IGI Global: Hershey,

PA, USA.
DOI: 10.4018/978-1-60566-766-9.

Jain A. & Dubes R. (1988). *Algorithms for Clustering Data*. Michigan State University: Prentice Hall.

Laslett, P. (1969). Size and Structure of the Household in England Over Three Centuries. *Population Studies,* 23(2), 199-223.
DOI:10.1080/00324728.1969.10405278

Laslett, P. (1972). Introduction. In: Laslett, P. with the assistance of Wall, R. (Eds.), *Household and family in past time. Comparative studies in the size and structure of the domestic group over the last three centuries in England, France, Serbia, Japan and colonial North America, with further materials from Western Europe*, (pp.1-89). Cambridge: Cambridge University Press.

Laslett, P. (1983). Family and household as work group and kin group: areas of traditional Europe com pared. In: Wall, R. in collaboration with Robin, J. and Laslett, P. (Eds.) *Family forms in historic Europe,* (pp 513-563). Cambridge: Cambridge University Press.

Laslett, P. (1985). Review. *Population and Development Review*, 11(3), 534-537.

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, Statistics, (pp 281–297). Berkeley: University of California Press.

Peres-Neto, P., Jackson, D. & Somers, K. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4), 974-997.
DOI: 10.1016/j.csda.2004.06.015

Ruggles, S. (2012). The Future of Historical Family Demography. *Annual Review of Sociology*, 38, 423-441.
DOI: 10.1145/annurev-soc-071811-145533.

Schürer, K. (1992). Variations in household structure in the late seventeenth century: towards a regional analysis. In: K. Schürer and T. Arkell, (Eds.) S*urveying the People. The interpretation and use of document sources for the study of population in the later seventeenth century*, (pp 253-278) Oxford: Leopard's Head.

Schürer, K. & Woollard, M. (2000). *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies, [original data producer(s)]. Colchester, Essex: UK Data Archive [distributor].
DOI: 10.5255/UKDA-SN-4177-1.

Schürer, K. & Woollard, M. (2002). *National Sample from the 1881 Census of Great Britain 5% Random Sample: working documentation version 1.1.* Colchester: University of Essex, Historical Censuses and Social Surveys Research Group.

Szołtysek, M., Gruber, S., Klüsener, S. & Goldstein, J. R. (2014). Spatial Variation in Household Structures in Nineteenth-Century Germany. *Population-E,* 69(1) 55-80.

Teitelbaum, M. S. (1984). *The British fertility decline: demographic transition in the crucible of the Industrial Revolution*. Princeton: Princeton University Press.

Wall, R. (1977). Regional and temporal variations in household structure from 1650. In: J. Hobcraft and P. Rees, (Eds.) *Regional demographic development,* 89-113) London.

Wall, R. (1982). Regional and temporal variations in the structure of the British household since 1851. In: T. Barker and M. Drake, (Eds.) *Population and society in Britain 1850-1980,* (pp 62-99). London.

Wall, R. (1983). The household: demographic and economic change in England, 1650-1970. In: Wall, R. in collaboration with Robin, J. and Laslett, P. (Eds.) *Family forms in historic Europe,* (pp 493-512). Cambridge: Cambridge University Press.

Woods, R. & Shelton, N. (1997). *An Atlas of Victorian Mortality.* Liverpool University Press: Liverpool.

Woods, R. (2000). *The Demography of Victorian England and Wales*. Cambridge: Cambridge University Press.

Wrigley, E. A. (1985). The fall of marital fertility in nineteenth-century France: Exemplar or exception? (Part II). *European Journal of Population*, 1, 141-177.
DOI: 10.1007/BF01796931

Wrigley, E. A. & Schofield, R. S. (1983). English population history from family reconstitution: summary results 1600-1799. *Population Studies,* 37, 157-184.
DOI: 10.1080/00324728.1983.10408745

Zinovyev A. (2000). ViDaExpert – multidimensional data visualization tool. Institute Curie, Paris.