Article:

Green, M.A., Vickers, D.W. and Dorling, D. (2014) A neighbourhood level mortality classification of England and Wales, 2006-2009. Health and Place, 30. 196 - 204. ISSN 1873-2054

https://doi.org/10.1016/j.healthplace.2014.09.011

A neighbourhood level mortality classification of England and Wales, 2006-2009

Mark A Green, Daniel Vickers, Danny Dorling

Abstract

The paper provides an overview of a neighbourhood level classification of mortality for England and Wales (2006-2009). Standardised mortality ratios for 63 causes of death were calculated for middle super output areas (weighted by prevalence). A k-means partitional method was used to classify the data. An eight cluster solution was found to best segment mortality patterns. Clusters mostly differentiated in terms of prevalence, however the importance of neurodegenerative diseases and causes related to unhealthy behaviours were important. The results describe a neighbourhood classification that can be an important tool to help inform policy development, resource allocation and targeting of services.

MeSH Key Words

Classification; Cluster Analysis; Mortality; Demography.

Highlights

- A multi-dimensional approach to conceptualising neighbourhoods
- Mortality patterns can be summarised by eight main groups
- Poverty was the most important factor in explaining the segmentation patterns
- The classification is useful at discriminating mortality patterns

Introduction

Although health is an individual level outcome, there is long-established evidence of geographical inequalities and patterns in health being found to be independent of individual-level explanatory factors (Diez Roux 2001; Thomas et al. 2010). Evidence has also shown that neighbourhoods can have an influence on individual health, through mechanisms such as the effects of living in areas of deprivation, geographical influences on social relations or the accessibility to services (Pickett & Pearl 2001; Riva et al. 2007; Diez Roux 2001). It is useful to examine the differences in health between places rather than just to assume that space is a passive factor that acts as a container for the existence of individuals who do not interact with people in the areas in which they live (Harris et al. 2005). Assuming any country to be spatially homogenous would restrict our understanding and ignore the geographical patterns that have been found to exist amongst the complex array of health patterns. It is through place that the underlying structure which differentiate the health and death of the population become most visible.

The importance of place provided to incentive for this study to develop novel approaches to measure and summarise the multiple geographies of health. Designing useful small area measurements of health is important for developing and targeting effective policies aimed at improving health. Previous approaches have focused on either geographical or socio-economic measures to group or define areas. These approaches allow the linkage to ecological attributes that help develop our understanding of how health outcomes occur. There has been less consideration of how the health characteristics of small geographical areas could be used to group areas to summarise their geographical patterning.

Geo-demographic classifications have sought to categorise areas in terms of the types of individuals that characterise them (Vickers & Rees 2007). Their popularity has seen the field develop into a multi-million pound industry with their resulting software tools being used in many commercial and public sector organisations (Harris et al. 2005). There has been less consideration of their application in the field of public health, beyond using geo-demographic classifications to identify population subgroups (Abbas et al. 2009; Nnoaham et al. 2010). This is despite calls for greater focus of such techniques within governmental policy and research in the field of public health (Department of Health 2005). The few that have tackled the field have been limited in scope. For example, Shelton and colleagues' (2006) area classification was conducted at a large geographical scale (parliamentary constituency), resulting in a loss of the wider variation that can be studied between smaller areas due to the coarse geographical areas used. The commercial company CACI have produced the commercial classification ‘HealthACORN’ as subsidiary of their main classification ‘ACORN’ however this is no longer available. There is currently no widely available area-based classification of mortality patterns at a small geographical scale.

Although geographical patterns in mortality have been reported (e.g. Shaw et al. 2008), they are often not disclosed for small geographies due to the small numbers involved when considering specific causes of death (resulting in confidentiality and outlier issues). This has restricted the application of area-based classifications in the field (Abbas et al. 2009). Furthermore with a wide range of causes of death to consider, common patterns and processes found within particular groups of causes can become lost limiting our capacity to understand and analyse such information. The area classification described here addresses these issues through summarising the main patterns into a series of groups and classifying areas into which group they fall (Harris et al. 2005). It allows the discovery of a hidden structure to the data that would be otherwise difficult to see (Everitt et al. 2001). Through describing the structure of the data in a simpler form that retains the most important information, our classification makes these complex geographies of health manageable which improves our understanding about patterns and processes.

Williams and colleagues (2004) argue, “The drive to tackle health inequalities and the move to localised policy making have increased interest in small area mortality data.” (p958). However, research has focused on analysing mortality causes independently of each other. An area classification allows a move away from one-dimensional approaches to a multi-dimensional understanding of the health of areas through summarising the main patterns across a range of variables. Knowing and understanding how causes of death vary and co-associate within an area is important for the effective targeting of policies, resources and location of services (Murray et al. 2006).

This paper details the construction of an area classification of mortality patterns in England and Wales (2006-2009). It draws from a rich database containing mortality data at a low geographical scale across all known causes of death. The results from the study will hopefully provide a useful tool for researchers and policy-makers.

Methods

Data

In England and Wales, it is a legal requirement to report every death. The Office for National Statistics (ONS) collects this information and collates every death into a database (Devis & Rooney 1999; Griffiths et al. 2005). Access to an anonymised version of the database was granted to the researchers by the ONS 'Microdata Release Panel' (December 2010). The database included records for every death registered between 1981 and 2009 with data on year of death, age at death, sex, cause of death and a geographical location identifier (postcode).

Cause of death is recorded using the International Classification of Diseases (ICD), which was developed to standardise mortality statistics between countries and is updated over time to account for medical advances (Rooney & Smith 2000; World Health Organisation 2004). It is filled in by the doctor who last treated the deceased and records the underlying cause of death (Devis & Rooney 1999; World Health Organisation 2004). ICD-9 is used for the years 1981-2000 and ICD-10 for 2001-2009. We chose to focus on the time period 2006-2009 to provide enough deaths per geographical unit (MSOA) to give stable estimates, whilst not being too wide temporally to lose accuracy. There were no missing data between 2006 and 2009 for individuals aged above zero (with 30% of deaths for age zero missing).

Data coding through the ICD-10 is based upon a hierarchical classification (Devis & Rooney 1999). Individual causes of mortality are grouped into broader ICD chapters which reflect their similarities. These ICD chapters are based on diseases of specific organs, pathology, aetiology, as well as more external causes or those related to specific time periods (Griffiths et al. 2005; World Health Organisation 2004). Within the ICD chapters, causes are grouped by type. At its lowest level, which gives the specific type or site of a particular cause, there are over 14,000 codes (World Health Organisation 2004). It is important to include a practical number of input variables that maintain the variation and detail in the data.

All deaths were included since no other study has carried an investigation using such small geographical areas. Little is known about how a large range of mortality variables are co-associated with each other. There is also less theoretical basis for just focusing on a few (and hence limiting our ability to describe areas; Voas & Williamson, 2001). Variable selection only included causes that contained at least 0.5 per cent of the total of deaths throughout the study period. This choice was due to statistical reasons, since it gives a figure greater than the number of areas (0.5% deaths was 9955 and there were 7194 areas). Therefore an even distribution would at least provide more than one death in each area, in line with recommendations from the cluster analysis literature (Everitt et al. 2001; Gordon 1999; Milligan & Cooper 1987).

Cause of death was coded to three digits using ICD-10. Those causes above the 0.5% threshold were then included in the model. The remaining codes were then combined into relevant categories based upon their ICD-10 groupings to fulfil the same criteria. However, this resulted in four variables that contained 27 ('Disease of the Eye and Adnexa'), 77 ('Disease of the Ear and Mastoid Process'), 160 ('Causes Related to Pregnancy and Childbirth') and 63 ('Conditions Originating in the Perinatal Period') deaths each. The variables were excluded due to their small numbers. Cases with the code U50 (n=2072) were not included as these were cases which have been sent to the coroner pending further investigation (Rooney & Smith 2000). ICD-10 data for data for individuals aged zero were combined into one variable for 'Infant Mortality' as 70 per cent of individuals aged zero had cause of death data missing. The 63 variables chosen are presented in Table 1.

(Table 1 here)

Individual deaths were aggregated using their postcodes to the Middle Super Output Area (MSOA) geographical scale. This was chosen as the geography is designed to be socially homogenous, the areas are similar in terms of population size (mean population size 7200) and their boundaries are designed to be relatively stable over time which is important for the application and dissemination of the classification (Office for National Statistics 2011). The mean number of deaths per MSOA was 275.

To control for the age and sex make-up of each MSOA, the mortality data were converted into Standardised Mortality Ratios (SMRs). An indirect approach was selected as it is more stable for dealing with small numbers. Population estimates by age and sex at the MSOA level were provided by the ONS. The SMRs were weighted by their prevalence to make them more representative of the actual structure of mortality patterns (rather than each variable being of equal importance).

Data were also collected from the 2011 Census and the ONS for MSOAs to help interpret the area classification. The number of communal establishments in 2011 (e.g. nursing homes, care homes) and the net migration rate for those aged over 65 (2008-2009) were included to explore the role of movements of the elderly, which have been shown to be important previously (Williams et al. 1995). Modelled estimates of the percentage of households in an area with an equivalised income of less than 60 per cent of the median income for England and Wales (2007-2008) was also available to explore the role of social disadvantage (Gregory 2009).

Statistical Analysis

Cluster analysis is a family of statistical methods used to group a diverse range of heterogeneous cases into a smaller number of more homogenous clusters (Gordon 1999). Rather than test hypotheses, cluster analysis seeks to analyse how cases (in this case; areas) are related. The analysis is exploratory, seeking to describe the underlying structure of the data to help discover new groups and identify future research directions (Everitt et al. 2001; Harris et al. 2005). Whilst a method of data reduction, it allows the simplification of complex data to help analyse the similarities and differences in the data that would otherwise be difficult to see.

A k-means partitional method was used to classify the data. The method operates through partitioning the data into a specified number of clusters (Gordon 1999). Cases are then individual relocated to the cluster centre that they lay nearest to and then the move is assessed to see if it improves the model. Cases are then reassigned until not further improvements can be gained (Everitt et al. 2001). Selecting a robust methodology is important in exploratory research, since the results should be due to variations in the data rather than the method used. The method was chosen as it is less influenced by outliers and quicker at processing larger data sets than traditional hierarchical methods (Gordon 1999; Everitt et al. 2001; Harris et al. 2005). Since only one partition of the data is sought, the method iteratively refines the partition to create an optimal solution which other methods do not offer. Euclidean distance was used to measure distance between objects (for assessing their degree of similarity).

The choice of seed points is an important issue involved with the k-means methodology. K-means begins by making an initial partition of the data into the specified number of groups (Everitt et al. 2001). How these initial groups are selected is important as the final solution may be dependent on them. A seed point may lead to the formation of a small specific cluster that does not accurately reflect the underlying structure of the data (Gordon 1999; Harris et al. 2005). Typically the seed points are randomly selected cases from the data, with group membership at the first iteration allocated based on similarity to them. However, this approach is sensitive to error as the selection of outliers can produce extreme clusters that do not reflect the main patterns (Milligan 1980). To reduce this influence of bias on the solution, hierarchical cluster analysis (Ward's method) was used to define the seed points. Krieger and Green (1999) note that this approach is more effective than using random sampling as it is less influenced by outliers. The centroids of the chosen solution were used as seed points for the k-means method to refine to find the optimal solution.

An issue with the method is that the number of groups to partition the data into needs to be specified a priori (Everitt et al. 2001; Gordon 1999). However, the correct number of groups in the data is not known. To inform this decision, a selection of statistics were used to choose the number of clusters that best represents the data. These included the 'C-index' (Hubert & Levin 1976) and Davies & Bouldin (1979) 'Validity Index' which both examine the similarity of clusters. These measures have both been shown to be effective at determining the correct number of clusters (Milligan & Cooper 1985). Ad hoc measures 'mean distance of cases to their cluster centre' (indicating compactness of clusters) and 'average cluster size' (showing if the clusters were evenly sized) were also used (Everitt et al. 2001). The measures were calculated by performing the analysis on a series of possible solutions between two and 16, as it was expected that the true solution would lie between these bounds (as well as being effective with the dissemination of the results).

Figure 1 presents the results of this initial process. The measures are designed so that they naturally improve as the number of clusters increase. Assessing the range of solutions should not be purely on the metric's value, rather it is important to select the solution which balances the extra detail offered by a larger number of clusters and the simplification of patterns gained from having a smaller number. The gradient of each measure was examined to identify the 'knee point' in the trend of increasing number of solutions. Where there is a large improvement in the model, followed by a flattening of the trend, this would indicate that further improvements in the number of clusters are not adding much more understanding to the classification (Milligan & Cooper 1985). Comparing the measures in Figure 1 suggested that an eight cluster solution was most appropriate for capturing the variations in the data as it performed most efficiently across each measure.

(Figure 1 here)

Statistical testing was employed to assess the stability of the resulting area classification. The approaches emphasize accuracy rather than precision. It is less important to assess if areas are correctly classified, but that the main relationships described by the clusters overall fit a stable and robust structure (Milligan 1996; Gordon 1999). Testing will demonstrate how useful the clusters are, since if they are found to be stable then this would suggest that they are effective at describing the structure of the data.

Blashfield and McIntyre's (1980) split-sample method was used as a replication analysis, as there were no alternative data sets for replication of the clusters (given that all mortality data for England and Wales was included). Through randomly splitting the data in half, we would expect that if the clusters were distinct then then they would be found in both samples. Similar clusters were found using the method, suggesting the persistence of the main patterns captured in the classification. Where there were changes in cluster membership, these were areas that moved to clusters of similar mortality profiles.

The influence of outliers on the solution was explored using Cheng and Milligan's (1996) framework through measuring outliers in relation to areas that lied far from their cluster centre (i.e. outliers in terms of cluster membership). Outliers were identified at ±1 and ±3 standard deviations, removed and the analysis re-run. There was little change in the results, with the clusters remaining broadly the same and changes in group membership between clusters of similar mortality profiles.

Finally, the sensitivity of the area classification to each input variable was conducted to examine what variables were driving the cluster formation (Milligan 1996). This was conducted through removing each variable individually, re-running the analysis and comparing the output to the original classification. Few variables had a large impact on the results, suggesting that they all had some useful contribution to the classification. The variables ‘Dementia’ and ‘Senility’ had a strong impact, however removing them from the analysis and re-running it resulted in little change in the classification showing that their influence was not problematic to the classification.

Results

The cluster centres (i.e. the mean characteristics of areas within each cluster) are presented in Table 2 to allow the understanding of what each cluster represents. The values were reconverted from their weighted values back to SMRs to improve their interpretation. A value of 100 is equivalent to the average for England and Wales for a particular cause, with a value above this representing the percentage increase in the mortality rate in a cluster compared to the national average (and vice versa). Life expectancy estimates at birth (split by sex) and sample size are also included. The cluster centres were converted from their weighted scores back to SMRs to aid interpretation. Each cluster has also been named with respect to its characteristics.

(Table 2 here)

Table 2 shows the varying mortality patterns between the clusters, with each capturing a different profile. The clusters ‘Poor Health Experiences’, ‘Poorest Health and Least

Desirable' and 'Poorest Neurodegenerative Health' contained areas with the worst mortality rates, differing by level of prevalence and the causes that dominated their profiles. 'Best Health and Most Desirable' and 'Good Health Areas' were the clusters with the lowest mortality rates, with the former being less prevalent. 'Average Mortality Profiles' contained areas with SMRs that fluctuated around 100 (apart from neurodegenerative diseases). 'Mixed Experiences' contained varied SMRs, with high prevalence of neurodegenerative diseases. Finally, the characteristics of the cluster 'The Middle' fell in-between those of the other clusters, with above average mortality rates.

Through exploring the range of values for each variable in Table 2, an assessment can be made regarding the impact of each variable in the area classification. A large difference in values shows a greater variation in patterns. The highest ranges were observed for those causes which display the greatest social and geographical patterning (e.g. 'Lung Disease Due to External Agents', 'Alcoholic Liver Disease', 'Chronic Lower Respiratory Diseases' etc) (Shaw et al. 2008). The neurodegenerative diseases 'Dementias' and 'Alzheimer's' were also important indicating their geographical clustering (although this was mostly driven by the clusters 'Poorest Health and Least Desirable' and 'Poorest Neurodegenerative Health'). Most of the cancer-related variables formed the variables with the lowest ranges, indicating that they are less distinctly patterned.

Life expectancy estimates mapped well onto the mortality profiles for each cluster, where those clusters which displayed worse mortality profiles had lower estimated life expectancy. The range of life expectancy values was 9.1 for males and 7.9 for females representing wide geographical inequalities. The gap between male and female life expectancy varied by cluster, being larger in those clusters with worse mortality profiles.

Figure 2 maps the spatial distribution of the clusters. Whilst it is difficult to see common geographical patterns, there are clear urban-rural divides, with the clusters with good mortality profiles more common in rural areas and vice versa. Also included is the map for London, which presents a distinct geographic pattern. The area classification follows the traditional East-West divide, with those clusters with poorer mortality profiles being more common in the East (Green 2012). There are a greater range of experiences (i.e. mortality profiles) in the more deprived East of London, suggesting a tailored approach is required to tackling areas with poor health. Similar patterns emerge in other urban areas as well (results not shown).

(Figure 2 here)

Table 3 presents three important factors that aid the interpretation of the clusters. Clusters which contained higher levels of household poverty were associated with poorer mortality profiles, showing the area classification to be partly capturing social divisions. The demographic factors of migration patterns of those aged over 65 and the location of communal establishments were also important in explaining the clusters 'Poorest Health and Least Desirable', 'Poorest Neurodegenerative Health' and 'Mixed Experiences'. Migration of the elderly into these clusters, possibly associated with the use of communal establishments, appears important in explaining them (especially for the latter two clusters, where the level of poverty is lower than expected given their mortality profiles).

(Table 3 here)

Table 4 compares the ranges of life expectancy estimates found between the clusters, to the ranges for other similar approaches to grouping areas into a similar number of categories for analysing the geographies of health. There is a greater range for life expectancy estimates found using the area classification than compared to other geographical and social approaches to grouping areas using a similar number of areas.

(Table 4 here)

Discussion

The paper gives an overview of the creation of a neighbourhood level classification of mortality for England and Wales. The tool segments England and Wales into eight main clusters of areas which describe the underlying structure of mortality patterns. Wide geographical inequalities in mortality continue to persist and therefore producing greater evidence (and a research tool) at small geographical scales to inform policy development is paramount to tackle them (particularly for resource allocation and the targeting of services) (Abbas et al. 2009; Harris et al. 2005; Williams et al. 2004). It moves the application of mortality statistics away from uni-dimensional measures towards a multi-dimensional approach for conceptualising the 'health' of an area. It is important to consider all causes of death since diseases do not operate in isolation and the area classification helps bring clarity to the complexities of mortality patterns for all causes of death (Everitt et al. 2001). A list of MSOAs and their respective cluster group for use in policy development and research can be found at Green (2013).

The main factor dividing England and Wales into the clusters appears to be levels of prevalence of overall mortality. However variations by cluster are not just dissimilar scales of the SMR scores, rather there are differences by cause as well. For example, 'Poorest Health and Least Desirable' and 'Poorest Neurodegenerative Health' are similar, containing high mortality rates. Yet the causes which dominate each cluster are different. With 'Poorest Health and Least Desirable', there are high rates for causes related to unhealthy behaviours (for example respiratory, digestive and some heart-related causes). 'Poorest Neurodegenerative Health' is not just a function of slightly lower rates, being characterised by higher rates for mental and nervous system related causes. This is important with regards policy implementation, as these differences require different approaches that otherwise may be missed using single mortality measures.

Social processes were also important, with the causes of death that are known to be more socially determined in their distributions being more influential in cluster formation. Poverty was a particularly important factor in explaining the clusters, showing the importance of such factors in understanding the underlying structure of mortality patterns (Shaw et al. 2008; Gregory 2009). Demographic factors were also important, particularly the role of the elderly gravitating towards similar areas at the end of life (Williams et al. 1995; Williams et al. 2004). The differences in these patterns should be explored in future research.

Estimates of life expectancy showed wide geographical inequalities by cluster. There are clear social injustice in that individuals living in particular areas (or ending up moving towards such areas) can expect to live on average nine or eight years longer (for males and females respectively). Examining the difference between life expectancy estimates between genders showed a greater gap in the clusters that were more deprived and experienced worse health. Males are more susceptible to social and spatial processes, reflecting the protective biological characteristics of females (Christensen et al., 2001).

The life expectancy estimates allowed an evaluation of the effectiveness of using the area classification as a tool for research. By considering the range of values, an assessment can be made regarding the discriminatory power of the area classification. The area classification offers greater discrimination of patterns than, for instance, equivalent means for segmenting England and Wales using a similar number of groups, including deprivation quintiles (Smith et al. 2010) and geographical regions (Office for National Statistics 2011). The area classification offers an alternative useful analytical tool for capturing detailed geographical information.

Limitations

There are several limitations to this study. The area classification represents a static classification. It presents a snapshot of the mortality patterns across England and Wales between 2006 and 2009. As a result, the classification is already out-of-date, remaining only directly applicable to that particular period (Harris et al. 2005). However changes in mortality patterns almost always occur slowly, driven by long term social processes (Rooney & Smith 2000). Whilst there may be changes in the cluster some areas lie in, it would not be expected that the patterns captured by the main clusters would have changed since its developed, meaning that any application would still be useful.

The process of running a cluster analysis contains subjective decisions that can affect the results (Everitt et al. 2001; Gordon 1999). This is because there are no set statistics which can objectively answer what precise decisions should be made at each stage of the analysis (Milligan & Cooper 1987). To minimise any issues, decisions made in the process (e.g. determining the number of clusters, testing the stability of the area classification) were informed by multiple factors to ensure that each decision was justified. Nevertheless there remains a degree of subjectivity in the outcome reached.

Communal establishments impacted upon the results. With their higher death rates, they can lead to bias in the data for the areas that they are located (Williams et al. 1995; Williams et al. 2004). Communal establishments were particularly important in explaining ‘Poorest Neurodegenerative Health’ and ‘Mixed Experiences’. However, not including these areas in the analysis would present a ‘false’ classification that does not describe mortality patterns for the whole of England and Wales and therefore the results will represent a group of areas that is less applicable to national policy makers. Instead, the influence of the geography of communal establishments should be viewed as helping to understand the geographical patterns of mortality through the area classification. These institutions reflect the changing nature of death, as their concentrated impact has become more prevalent with an ageing population (Williams et al. 2004). To address this issue, further research could look to adapt Bayesian modelling of the SMRs to minimise the effect of extreme data points.

Acknowledgements

Many thanks for the trust placed by Alan Farrell in the project. This research was carried out whilst Mark A Green was funded by an ESRC +3 studentship (ES/I023224/1) to explore mortality patterns in England and Wales. Thanks also to funding from the University of Sheffield and Oxford University.

References

Abbas, J., Ojo, A., Orange, S., 2009. Geodeomographics - a tool for health intelligence. Public Health 123, e35–39.

Blashfield, R., McIntyre, R., 1980. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioural Research 15, 225–238.

Cheng, R., Milligan, G., 1996. Measuring the Influence of Individual Data Points in a Cluster Analysis. Journal of Classification 13, 315–335.

Davies, D., Bouldin, D., 1979. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1, 224–227.

Department of Health, 2005. Choosing Health: Making health choices easier. Available at: http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/Browsable/DH_4097491 [Accessed February 14, 2014].

Devis, T., Rooney, C., 1999. Death certification and the epidemiologist. Health Statistics Quarterly 1, 21–33.

Diez Roux, A. V, 2001. Investigating neighborhood and area effects on health. American Journal of Public Health 91, 1783–1789.

Everitt, B., Sabine, L., Leese, M., 2001. Cluster Analysis. Arnold, London.

Gordon, A.D., 1999. Classification. Chapman and Hall, London.

Green, M.A., 2012. Mapping Inequality in London: A Different Approach. The Cartographic Journal 49, 247–255.

Green, M.A., 2013. Death in England and Wales: Using a classificatory approach for researching mortality. PhD Thesis, University of Sheffield. Available at: http://etheses.whiterose.ac.uk/5631/ [Accessed August 1, 2014].

Gregory, I.N., 2009. Comparisons between geographies of mortality and deprivation from the 1900s and 2001: spatial analysis of census and mortality statistics. BMJ 339, b3454.

Griffiths, C., Rooney, C., Brock, A., 2005. Leading causes of death in England and Wales - how should we group causes? Health Statistics Quarterly 28, 6–17.

Harris, R., Sleight, P., Webber, R., 2005. Geodemographics, GIS and Neighbourhood Targeting. Wiley, Chichester.

Hubert, L., Levin, J., 1976. A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin 83, 1072–1080.

Krieger, A., Green, P., 1999. A cautionary note on using internal-cross validation to select the number of clusters. Psychometrika 64, 341–353.

Milligan, G., 1980. An estimation of the effects of six types of error perturbation on fifteen clustering algorithims. Psychometrika 45, 325–342.

Milligan, G., 1996. Clustering Validation: Results and Implications for Applied Analyses. In Arabie, P., Hubert, L., De Soete, G. (Eds), Clustering and Classification. World Scientific, London, pp. 341–375.

Milligan, G., Cooper, M., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.

Milligan, G., Cooper, M., 1987. Methodological Review: Clustering Methods. Applied Pscyhological Measurement 11, 329–354.

Murray, C.J., Kulkarni, S., Michaud, C., Tomijima, N., Bulzacchelli, M., Iandiorio, T., Ezzati, M. 2006. Eight Americas: Investigating Mortality Disparities across Race, Counties, and Race-Counties in the United States. PLoS medicine 3, e260.

Nnoaham, K., Frater, A., Roderick, P., Moon, G., Halloran, S. 2010. Do geodemographic typologies explain variatiosn in uptake in colorectal cancer screening? An assessment using routine screening data in the south of England. Journal of Public Health 32, 572–581.

Office for National Statistics, 2011. Life expectancy at birth and at age 65 by local areas in the United Kingdom, 1004-06 to 2008-10. Available at: http://www.ons.gov.uk/ons/rel/subnational-health4/life-expec-at-birth-age-65/2004-06-to-2008-10/statistical-bulletin.html [Accessed February 14, 2014].

Pickett, K.E., Pearl, M., 2001. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. Journal of Epidemiology & Community Health 55, 111–122.

Riva, M., Gauvin, L., Barnett, T., 2007. Toward the next generation of research into small area effects in health a synthesis of multilevel investigations published since July 1998. Journal of Epidemiology & Community Health 61, 853–861.

Rooney, C., Smith, S., 2000. Implementation of ICD-10 for mortality data in England and Wales from January 2001. Health Statistics Quarterly 8, 41–50.

Shaw, M., Thomas, B., Davey Smith, G., Dorling, D. 2008. The grim reaper’s road atlas: an atlas of mortality in Britain. Policy Press, Bristol.

Shelton, N., Birkin, M., Dorling, D., 2006. Where not to live: a geo-demographic classification of mortality for England and Wales, 1981-2000. Health & Place 12, 557–569.

Smith, M., Olatunde, O., White, C., 2010. Inequalities in disability-free life expectancy by area deprivation: England, 2001-04 and 2005-08. Health Statistics Quarterly 48, 1–22.

Thomas, B., Dorling, D., Smith, G.D., 2010. Inequalities in premature mortality in Britain: observational study from 1921 to 2007. BMJ 341, c3639.

Vickers, D.W., Rees, P.H., 2007. Creating the UK National Statistics 2001 output area classification. Journal of the Royal Statistical Society: Series A 170, 379–403.

Voas, D., Williamson, P., 2001. The diversity of diversity: a critique of geodemographic classification. Area 33, 63–76.

Williams, E., Dinsdale, H., Eayres, D., Tahzib, F. 2004. Impact of nursing home deaths on life expectancy calculations in small areas. Journal of Epidemiology & Community Health 58, 958–962.

Williams, E., Scott, C., Scott, S., 1995. Using mortality data to describe geographic variations in health status at sub-district level. Public Health 109, 67–73.

World Health Organisation, 2004. International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Volume 2. Available at: http://www.who.int/classifications/icd/ICD-10_2nd_ed_volume2.pdf [Accessed February 14, 2014].

Tables

Cause of death	ICD10 Codes	Number of deaths	Per cent of total deaths
Cancers			
Cancer of the Gullet	C15	26098	1.3%
Stomach Cancer	C16	17994	0.9%
Colon Cancer	C18	35393	1.8%
Rectum Cancer	C20	14328	0.7%
Liver Cancer	C22	11507	0.6%
Pancreatic Cancer	C25	27390	1.4%
Lung Cancer	C34	118885	6.0%
Breast Cancer	C50	42824	2.2%
Ovarian Cancer	C56	14909	0.8%
Prostate Cancer	C61	36811	1.9%
Kidney Cancer	C64	12252	0.6%
Bladder Cancer	C67	17556	0.9%
Cancer of the Brain	C71	12738	0.6%
Leukaemia's	C91-95	15588	0.8%
Other Lymphatic Cancers	C81-90, 96	26748	1.3%
Other Cancers	Rest of C's, D00-48	127490	6.4%
Mental and Nervous System			
Dementias	F00-03	61315	3.1%
Other Mental and Behavioural Disorders	F04-99	6362	0.3%
Parkinson's Diseases	G20-22	18040	0.9%
Alzheimer’s	G30	23015	1.2%
Other Diseases of the Nervous System	G00-13, 23-26, 31-99	24893	1.3%
Respiratory			
Pneumonia	J12-18	112279	5.6%
Chronic Lower Respiratory Diseases	J40-47	103135	5.2%
Lung Diseases due to External Agents	J60-70	12031	0.6%
Other Diseases of the Respiratory System	J00-11, 20-39, 80-99	48838	2.5%
Heart			
Hyperintensive Diseases	I10-15	17266	0.9%
Acute Myocardial Infarction	I21	120378	6.1%
Chronic Ischaemic Heart Disease	I25	188496	9.5%
Pulmonary Heart Disease and Diseases of Pulmonary Circulation	I26-28	13852	0.7%
Atrial Fibrillation and Flutter	I48	12878	0.7%
Heart Failure	I50	33275	1.8%
Other Heart Diseases	I00-09, 20, 22-24, 30-47, 49, 51-2	40630	2.0%
Intracerebral Haemorrhage	I61	17878	0.9%
Cerebral Infarction	I63	17917	0.9%
Stroke	I64	88897	4.5%
Other Cerebrovascular Diseases	I60, 62, 65-69	59867	3.0%
Aortic Aneurysm and Dissection	I71	29921	1.5%

Diseases of Veins, Lymphatic Vessels and Lymph Nodes, Not Elsewhere Classified	I80-89	15489	0.8%
Other Circulatory Diseases	I70, 72-79, 95-99	13598	0.7%
Digestive System			
Ulcers	K25-28	11582	0.6%
Vascular Disorders of the Intestine	K55	9455	0.5%
Other Diseases of Intestines	K56-63	19629	1.0%
Alcoholic Liver Disease	K70	18270	0.9%
Other Liver Diseases	K71-76	11166	0.6%
Diseases of Gallbladder, Binary Tract and Pancreas	K80-86	10666	0.5%
Other Diseases of the Digestive System	K00-24, 29-54, 64-67, 90-93	21311	1.1%
Other			
Infant Mortality	All cases where age = 0	12909	0.7%
Septicaemia	A40-41	8929	0.5%
Other Infectious and Parasitic Diseases	A00-39, 42-B99	18546	0.9%
Diseases of the Blood	D50-89	3926	0.2%
Diabetes Mellitus	E10-14	21649	1.1%
Other Endocrine, Nutritional and Metabolic Diseases	E00-07, 15-90	6993	0.4%
Diseases of the Skin and Subcutaneous Tissue	L00-99	7359	0.4%
Diseases of the Musculoskeletal System and Connective Tissue	M00-99	16936	0.9%
Renal Failure	N17-19	11903	0.6%
Other Diseases of the Genitourinary System	N00-16, 20-99	33912	1.7%
Congenital Malformations, Deformation and Chromosomal Abnormalities	Q00-99	3887	0.2%
Senility	R54	35176	1.8%
Other Symptoms, Signs and Abnormal Findings	R00-53, 55-99	5872	0.3%
Falls	W00-19	12801	0.6%
Other Accidents	V01-99, W20-X59	30352	1.5%
Intentional Self-Harm	X60-84	12361	0.6%
Other External Causes	X85-Y98	6632	0.3%

Table 1: Variables selected to be included in the cluster analysis.

Cause of death	Clusters							
	Best Health and Most Desirable	Average Mortality Profiles	Good Health Areas	The Middle	Poor Health Experiences	Poorest Health and Least Desirable	Poorest Neuro-degnerative Health	Mixed Experiences
Cancers								
Cancer of the Gullet	84	106	96	110	123	130	98	96
Stomach Cancer	75	109	91	124	146	160	125	90
Colon Cancer	93	98	98	107	104	114	105	100
Rectum Cancer	87	100	95	107	122	132	114	97
Liver Cancer	84	106	92	123	137	161	123	97
Pancreatic Cancer	96	102	99	106	107	116	98	97
Lung Cancer	68	113	83	127	173	190	124	83
Breast Cancer	95	97	100	100	101	106	115	104
Ovarian Cancer	100	97	102	96	99	93	97	102
Prostate Cancer	96	96	100	99	100	102	111	106
Kidney Cancer	89	99	95	107	111	130	112	104
Bladder Cancer	84	105	96	111	112	135	115	99
Cancer of the Brain	102	94	98	94	94	95	105	106
Leukaemia's	99	100	97	100	107	101	99	98
Other Lymphatic Cancers	98	100	99	102	100	111	101	98
Other Cancers	84	105	94	112	123	132	111	96
Mental and Nervous System								
Dementias	57	60	63	87	78	170	349	159
Other Mental and Behavioural Disorders	56	96	72	138	159	217	147	88
Parkinson's Diseases	83	71	83	87	70	130	198	145
Alzheimer’s	66	62	71	88	79	151	284	157
Other Diseases of the Nervous System	84	91	90	102	103	139	141	115

Respiratory								
Pneumonia	72	90	85	117	112	165	167	122
Chronic Lower Respiratory Diseases	61	109	79	134	176	214	138	90
Lung Diseases due to External Agents	72	97	83	111	141	195	162	100
Other Diseases of the Respiratory System	73	91	83	107	115	160	176	120
Heart								
Hyperintensive Diseases	84	107	91	115	119	141	131	108
Acute Myocardial Infarction	82	132	82	112	169	180	136	102
Chronic Ischaemic Heart Disease	67	85	107	148	114	169	128	96
Pulmonary Heart Disease and Diseases of Pulmonary Circulation	84	104	97	121	116	134	121	97
Atrial Fibrillation and Flutter	84	96	90	101	113	120	128	114
Heart Failure	84	100	92	103	118	138	134	111
Other Heart Diseases	92	95	102	106	103	113	103	103
Intracerebral Haemorrhage	90	99	97	112	118	127	110	99
Cerebral Infarction	76	90	89	110	113	143	148	114
Stroke	76	92	83	99	106	149	173	128
Other Cerebrovascular Diseases	70	71	82	100	95	154	210	134
Aortic Aneurysm and Dissection	86	104	97	116	114	116	111	94
Diseases of Veins, Lymphatic Vessels and Lymph Nodes, Not Elsewhere Classified	77	98	99	129	116	129	137	101
Other Circulatory Diseases	74	94	83	107	122	167	146	112
Digestive System								
Ulcers	79	105	90	124	132	158	124	94
Vascular Disorders of the Intestine	73	101	92	124	140	160	133	98
Other Diseases of Intestines	82	100	92	118	117	135	132	105
Alcoholic Liver Disease	58	105	76	143	182	241	140	81
Other Liver Diseases	71	114	85	131	153	188	131	87

Diseases of Gallbladder, Binary Tract and Pancreas	72	104	95	124	145	173	126	96
Other Diseases of the Digestive System	81	102	88	112	125	143	135	102
Other								
Infant Mortality	75	103	85	109	123	145	110	84
Septicaemia	78	100	88	119	130	166	124	103
Other Infectious and Parasitic Diseases	83	117	90	118	139	152	128	100
Diseases of the Blood	85	110	84	108	130	143	123	111
Diabetes Mellitus	66	105	82	117	132	177	178	117
Other Endocrine, Nutritional and Metabolic Diseases	77	101	93	110	124	150	152	96
Diseases of the Skin and Subcutaneous Tissue	79	103	88	111	117	134	154	112
Diseases of the Musculoskeletal System and Connective Tissue	86	100	94	101	109	124	133	107
Renal Failure	80	104	87	105	124	147	149	114
Other Diseases of the Genitourinary System	76	97	85	117	109	145	169	118
Congenital Malformations, Deformation and Chromosomal Abnormalities	80	105	97	105	126	142	132	105
Senility	77	62	71	68	75	148	191	156
Other Symptoms, Signs and Abnormal Findings	65	106	78	136	160	197	149	81
Falls	76	92	91	128	133	178	122	96
Other Accidents	83	96	93	109	118	140	116	106
Intentional Self-Harm	86	100	94	119	117	139	101	94
Other External Causes	69	106	85	123	127	158	124	92
Male life expectancy	81.3	77.9	79.4	75.9	74.5	72.2	75	78.5
Females life expectancy	85.1	82.5	83.5	80.6	79.7	77.2	78.6	81.7
Sample size	1562	1149	1309	854	656	296	322	1046

Table 2: Cluster centres of the neighbourhood classification of mortality for England and Wales, 2006-2009 (reported as standardised mortality ratios for each cause of death, also including life expectancy and sample size).

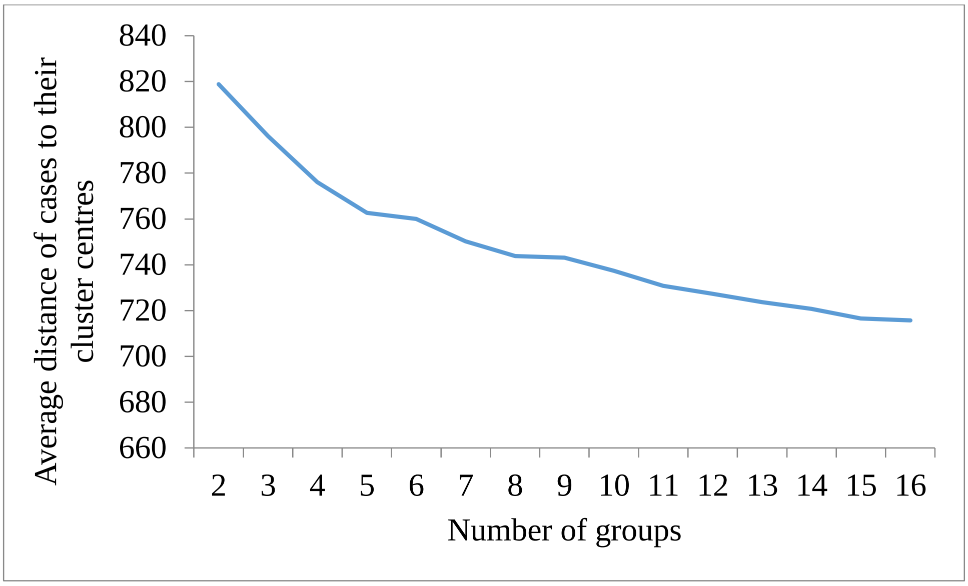
Cluster	Per cent of households in poverty	Net change of those aged 65 and over	Mean number of communal establishments
Best Health and Most Desirable	15.18	-4.28	7.6
Average Mortality Profiles	23.20	-6.78	6.5
Good Health Areas	19.12	-3.02	7.9
The Middle	27.09	-5.49	8.3
Poor Health Experiences	30.43	-7.33	5.6
Poorest Health and Least Desirable	31.97	1.77	11.5
Poorest Neurodegenerative Health	24.92	17.81	11.0
Mixed Experiences	18.32	10.44	10.6
Total	21.56	-1.49	8.2

Table 3: Explanatory factors of the clusters.

Measure	Range for males	Range for females	No. of areas	Source
Deprivation Quintiles	8	5.6	5	Smith et al. 2010
Governmental Office Regions	2.8	2.7	10	Office for National Statistics 2011
Mortality Classification	9.1	7.9	8	Table 2

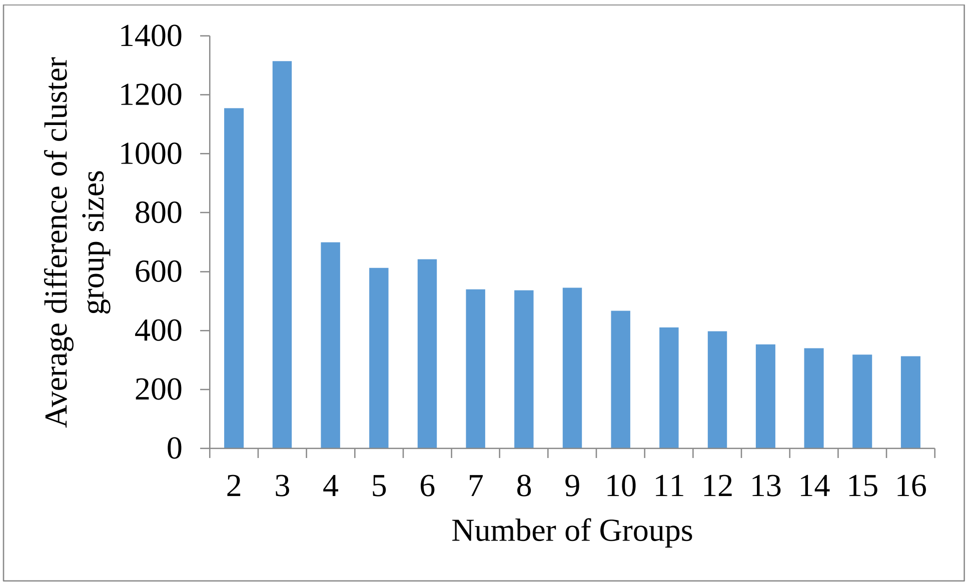
Table 4: Variation in life expectancy estimates by approaches to grouping areas together.

Figures



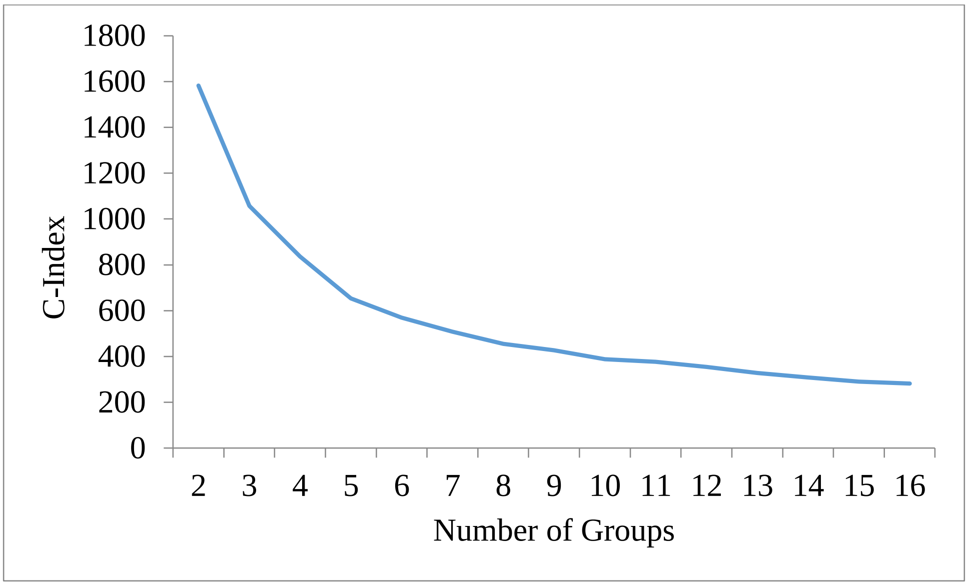
Average distance of cases to their cluster centres
840
820
800
780
760
740
720
700
680
660
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of groups

(a)



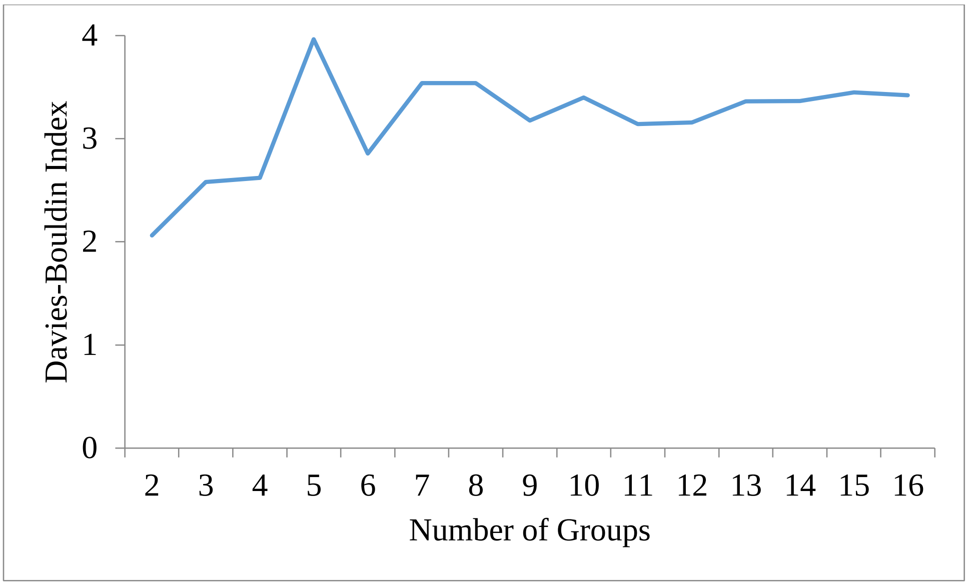
Average difference of cluster group sizes
1400
1200
1000
800
600
400
200
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of Groups

(b)



C-Index
1800
1600
1400
1200
1000
800
600
400
200
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of Groups

(c)



(d)

Figure 1: Assessment of cluster solutions through measures of cluster structure: (a) average distance of cases to their cluster centres; (b) mean difference in cluster size from expected size for evenly sized clusters (i.e. sample size divided by number of clusters); (c) C-index; (d) Davies-Bouldin Validity Index.

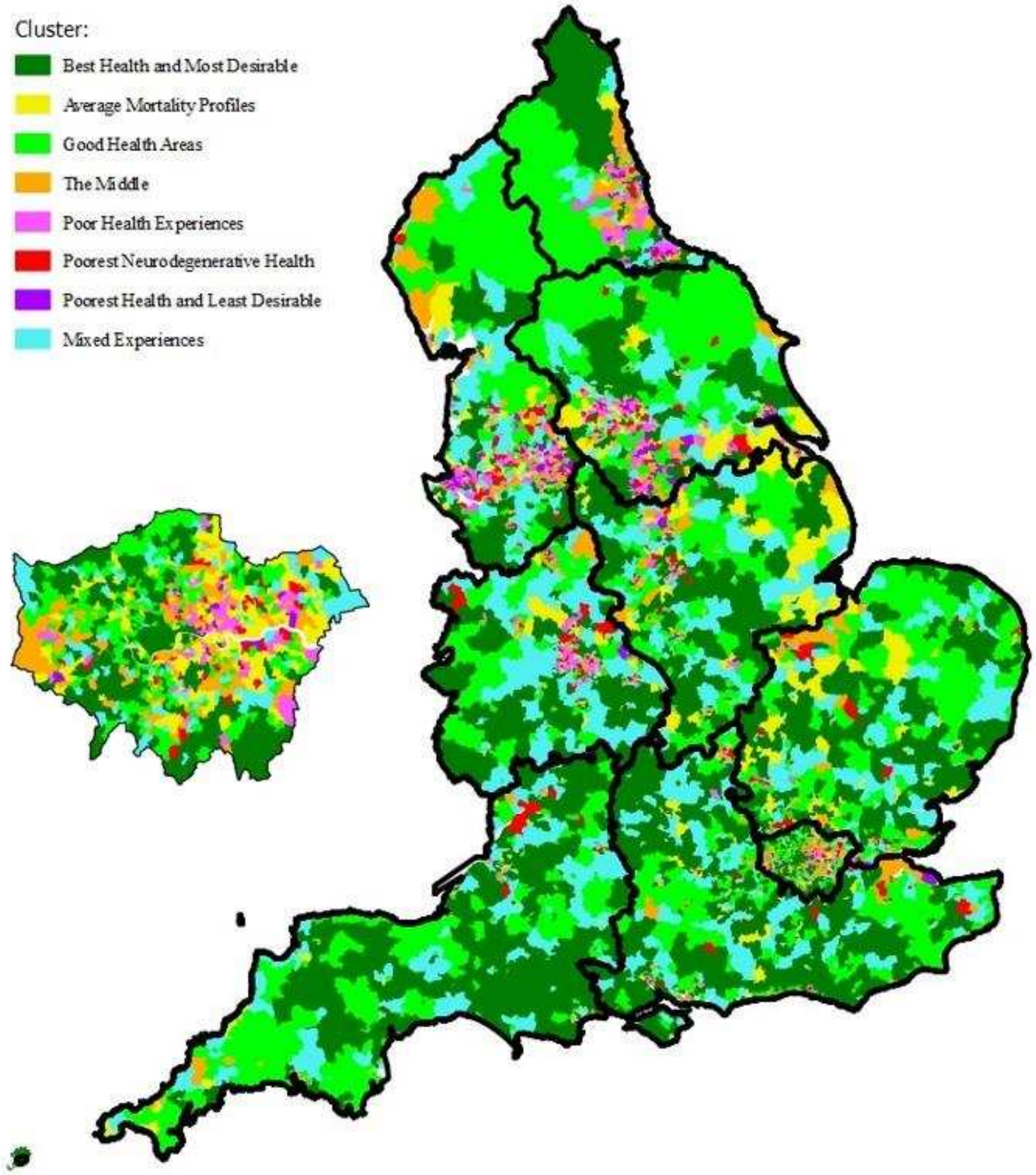

Cluster:
Best Health and Most Desirable
Average Mortality Profiles
Good Health Areas
The Middle
Poor Health Experiences
Poorest Neurodegenerative Health
Poorest Health and Least Desirable
Mixed Experiences

Figure 2: The geographical distribution of the area classification in England and Wales (with London inset).