What Types of Data Are Available for Mortality Data

Andrew J.G. Cairns

Heriot-Watt University, Edinburgh

Director, Actuarial Research Centre, IFoA

Joint work with Jie Wen and Torsten Kleinow

IFoA ARC Workshop
2 December 2019

The views expressed in this presentation are those of invited contributors and not necessarily those of the Institute and Faculty of Actuaries (IFoA). The IFoA does not endorse any of the views stated, nor any claims or representations made in this presentation and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this presentation.

The information and expressions of opinion contained in this presentation are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this presentation be reproduced without the written permission of the IFoA.

Actuarial
Research Centre
Institute and Faculty
of Actuaries

Outline

- Background
- Data – England male & female mortality
  - All-cause mortality data
  - Cause-of-death data
  - Predictive variables

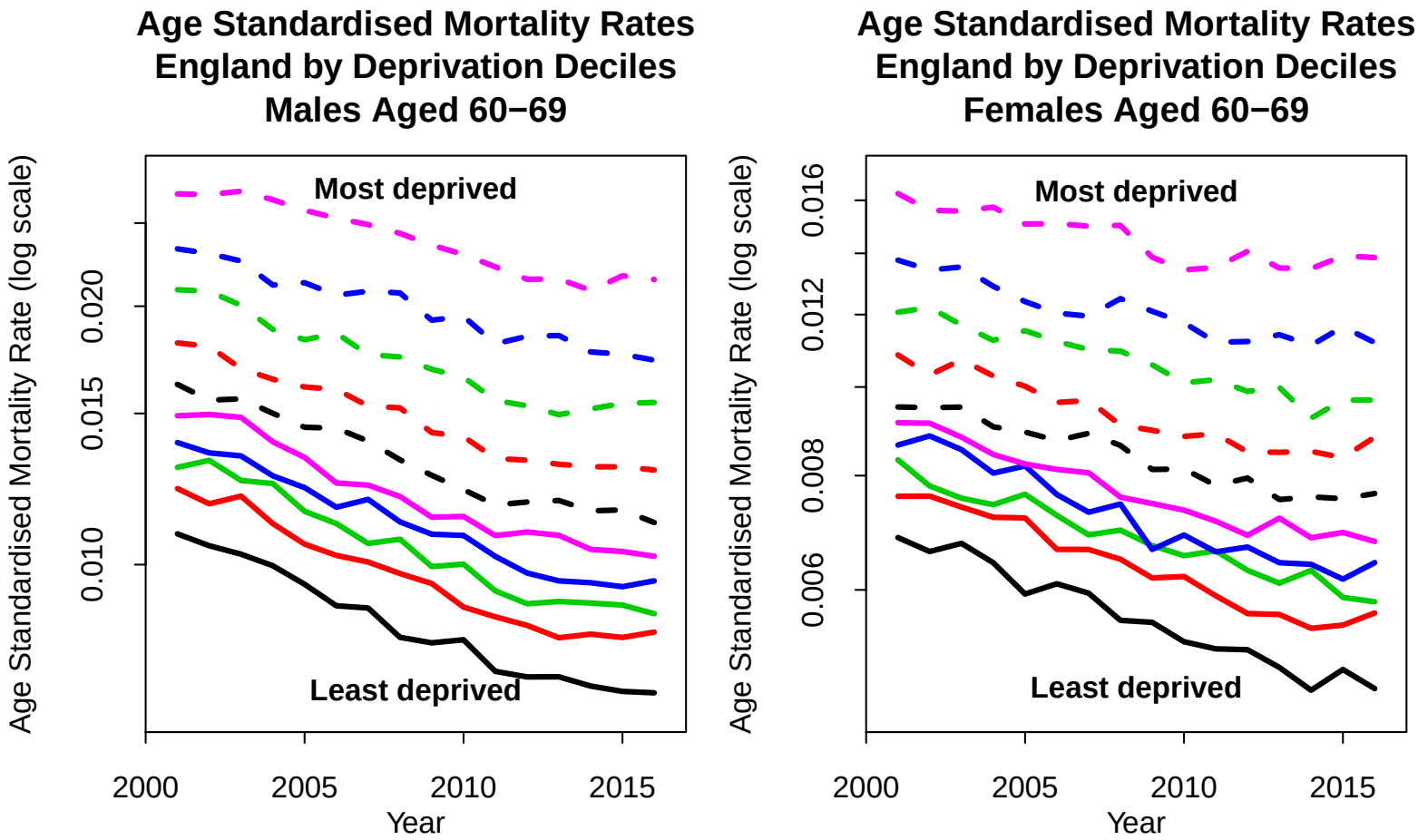
Background

- Focus for this workshop:
  male and female mortality in England
- Stylised facts:
  - Mortality varies by socio-economic group
  - Mortality varies by region

Actuarial Research Centre
Institute and Faculty of Actuaries

Socio-Economic Differences in Mortality: England

England: mortality by *deprivation*



Background: Variation By Region

North East
North West
Yorkshire & Humber
East Midlands
West Midlands
East of England
London
South East
South West

Not in dataset:
Scotland, Wales,
Northern Ireland

Background: Relative mortality by region

England Variation by region (males 60-69)	
North East	118%
North West	116%
Yorkshire and The Humber	107%
East Midlands	98%
West Midlands	105%
East	88%
London	105%
South East	89%
South West	87%

Values show actual deaths (ages 60-69) by region as a percentage of expected deaths using national age-specific mortality

Regional variation $<$ variation by income deprivation

Background

- Mortality varies by socio-economic group
- Mortality in the north (and in big cities) is higher than mortality elsewhere
- *How much of this can be explained by underlying socio-economic differences?*
- *And how much variation is geographical?*

  E.g. due to higher or lower levels of smoking than national levels by socio-economic group.

Actuarial Research Centre
Institute and Faculty of Actuaries

How much data?

- 'Small' data
  Deaths and exposures (population) for e.g. the national population (years $\times$ ages)
- 'Big' data
  E.g. data at the level of the individual including predictive variables; frequently updated
- 'Medium' data
  Data for many small geographical areas including area-specific predictive variables

Actuarial Research Centre
Institute and Faculty of Actuaries

Data: LSOA's

- England only
- Lower Layer Super Output Areas: LSOA's
- $L = 32,844$ small geographical areas
- Socio-economically homogeneous
- Average size ≈ 1600 persons
- LSOA's $i = 1, \ldots, L$,
  single years (t =2001-2016), single ages, x,
  gender, g:
  - Deaths: $D(g, i, t, x)$
  - Exposures: $E(g, i, t, x)$ (population)

Data: LSOA's (cont.)

- About 90% of the $D(g, i, t, x)$ are zero!
- About 6% of the $E(g, i, t, x)$ are zero.
- Exposures are *estimated* from census data at the LSOA level

  and returned as integers
- 0.6% of the (g, i, t, x) cells for ages 40-89 have $D(g, i, t, x) > 0$ but $E(g, i, t, x) = 0$!

  $\Rightarrow$ a problem unless data are aggregated

$$\text{e.g.} \sum_{t=t_0}^{t_1}, \quad \sum_{x=x_0}^{x_1}, \quad \text{or} \quad \sum_{t=t_0}^{t_1} \sum_{x=x_0}^{x_1}$$

or you have a model for errors in the $E(g, i, t, x)$.

Predictive variables by LSOA

- Indices of Deprivation (2015) (single scores per LSOA)
  - income deprivation (benefits)
  - employment deprivation (unemployment)
  - education deprivation
  - crime
  - barriers to housing and services
    - geographical barriers (distance to services)
    - wider barriers (overcrowding; homelessness; affordability)
  - living environment (housing quality; unmodernised; air quality)
- Educational attainment (levels $\times$ age groups)
- Occupation groups (types $\times$ age groups)
- Average weekly income
- Average number of bedrooms
- # people in care homes with/without nursing

Predictive variables by LSOA (cont.)

- Proportion UK born
- Country of birth
- Religion
- Ethnic group
- Urban/rural classification (categorical)
- Lookup:
  Postcode $\rightarrow$ OA $\rightarrow$ LSOA $\rightarrow$ MSOA $\rightarrow$ local authority $\rightarrow$ region

Predictive variables by LSOA (cont.)

1. LSOA index.
   - LSOA codes are of the form "E010xxxxx" where the LSOA index *xxxxx* ranges from 00001 to 33768.
   - Only 32,844 indexes are currently in use and, therefore, some codes are missing. These are codes that would have been used previously. However, if an LSOA has grown substantially, then it would be split, the old LSOA code deleted, and the two new LSOAs given new codes not yet used. And some LSOAs have shrunk and will have been merged and allocated a new index.
2. Lookup:
   Postcode $\rightarrow$ OA $\rightarrow$ LSOA $\rightarrow$ MSOA $\rightarrow$ local authority $\rightarrow$ region

Predictive variables by LSOA (cont.)

3. The Index of Multiple Deprivation (IMD)

Official composite measure of *relative deprivation* in England, with a single value for each LSOA. A higher value indicates a higher level of deprivation. The IMD has seven domains:

- income deprivation;
- employment deprivation;
- education, skills and training;
- health deprivation and disability;
- crime;
- barriers to housing and services;
- living environment.

Some of these have further sub-domains (which we discuss below)

that we consider to be useful to refine predictions of mortality.

Predictive variables by LSOA (cont.)

4. Income deprivation (a domain of the Index of Multiple Deprivation (IMD)):
   - this measures the proportion of the population in each LSOA who are receiving benefits from the state because they are on a low income;
   - the data are in a vector of length 32,844: one entry for each LSOA;
   - sub-domains include *income deprivation affecting older people*, which measures income deprivation amongst people aged 60 and older.

Predictive variables by LSOA (cont.)

5. Employment deprivation (a domain of the IMD)
   - this measures the proportion of the *working* population in each LSOA who are unemployed;
   - the data are in a vector of length 32,844: one entry for each LSOA corresponding to the vector of 5-digit LSOA codes above.

Predictive variables by LSOA (cont.)

6. Living environment deprivation (a domain of the IMD)
   - this measures the quality of the living environment (indoors and outdoors);
   - indoors: (poor) quality of housing;
   - outdoors: e.g. (poor) air quality and traffic accidents;
   - the data are in a vector of length 32,844: one entry for each LSOA.

Predictive variables by LSOA (cont.)

7. Barriers to housing and services (a domain of the IMD)
   - measures a number of different things: 'wider barriers' and 'geographical barriers';
   - wider barriers includes overcrowding in households, homelessness and affordability of housing;
   - geographical barriers measures distance to key services;
   - a higher value for geographical barriers $\Rightarrow$ more 'deprived',
     BUT
     might also imply lower mortality;
     e.g. greater distances to services might indicate that the LSOA is more affluent or rural with housing more spaced out;

Predictive variables by LSOA (cont.)

in fact, the geographical barriers variable is negatively correlated with income deprivation;

- the data are in a vector of length 32,844: one entry for each LSOA;
- data are available separately for wider barriers and geographical barriers.

8. Crime
   - Measures the risk of personal and material victimisation at local level

Predictive variables by LSOA (cont.)

Predictive variables that are not part of the IMD:

1. Average number of bedrooms
   - this measures the average number of bedrooms per household in the LSOA
   - the data vector has been standardised to a $N(0,1)$ distribution;
   - in contrast to the deprivation indices, a high value (more bedrooms) is likely to be associated with lower mortality;
   - the data are in a vector of length 32,844: one entry for each LSOA.

Predictive variables by LSOA (cont.)

2. Highest level of qualification:
   - this gives the *proportion* of a particular group within the LSOA who have attained a particular level of education
   - data are in the form of a 3-dimensional array for males and females combined
     lsoa x age-group x education level $(32,844 \times 6 \times 8)$
   - 6 age groups: All; 16 to 24; 25 to 34; 35 to 49; 50 to 64; 65 plus;
   - 8 education groups:
     1. All categories: Highest level of qualification
     2. No qualifications
     3. Level 1 qualifications (up to low grade GCSE's)
     4. Level 2 qualifications (higher grade GCSE's)
     5. Apprenticeship

Predictive variables by LSOA (cont.)

6. Level 3 qualifications (A-level)
7. Level 4 qualifications and above (Some college/university qualification, BSc, MSc, ...)
8. Other qualifications

- `www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels`
- you can use the education data to construct one or more vectors of predictive variables: e.g.
  - the proportion of people in the LSOA aged 50-64, who have no qualification or level 1 only;
  - an average level of educational attainment in a particular age group;

Predictive variables by LSOA (cont.)

3. Occupation group proportions
   - gives the proportion of a particular group within the LSOA who have a particular type of occupation
   - data are in the form of a 4-dimensional array gender x lsoa x age-group x occupation group $(2 \times 32,844 \times 14 \times 9)$
   - 14 age groups: All; 16-19; 20-24; 25-29; 30-34; 35-39; 40-44; 45-49; 50-54; 55-59; 60-64; 65-69; 70-74; 35-64
   - most age groups are small, so there will be a lot of sampling variation, weakening their predictive ability. This is less of a problem for the 35-64 age group.

Predictive variables by LSOA (cont.)

- 9 occupation groups
  1. Higher managerial, administrative and professional occupations
  2. Lower managerial administrative and professional occupations
  3. Intermediate occupations
  4. Small employers and own account workers
  5. Lower supervisory and technical occupations
  6. Semi-routine occupations
  7. Routine occupations
  8. Never worked, long-term unemployed and full-time students
  9. Total
- you can use the occupation data to construct one or more vectors of predictive variables

Predictive variables by LSOA (cont.)

4. Urban-Rural Classification
   1. Conurbation: non London
   2. City or town
   3. Rural town
   4. Rural village and dispersed
   5. Conurbation: London
   - the data are in a vector of length 32,844: one entry for each LSOA.

Predictive variables by LSOA (cont.)

5. Region
   1. North East
   2. North West
   3. Yorkshire and Humber
   4. East Midlands
   5. West Midlands
   6. East
   7. London
   8. South East
   9. South West
   - the data are in a vector of length 32,844: one entry for each LSOA.

Predictive variables by LSOA (cont.)

6. Communal establishments (own commissioned dataset)
   - This element of the data (a user-requested dataset from the ONS) record the number of persons in each LSOA in a communal establishment at the time of the 2011 census.
   - The data count the number of persons, $C(i, g, y, \tau)$ where
     - i is the LSOA index;
     - g is gender;
     - y is the age group 0-59, and 60+;
     - τ is the type of communal establishment:

     1 Care home: Private or local authority, with nursing;
     2 Care home: Private or local authority, without nursing;
     3 Remainder of medical and care establishments;
     4 Other communal establishments.

Predictive variables by LSOA (cont.)

7. Proportion of the population that are UK born
8. Proportions of the population in different ethnic groups
   (32 overlapping options)
9. Proportions of the population in different religious groups
   (9 options)
10. Country of birth
11. Average weekly income
12. Proportion of the population working more than 49 hours per week

Predictive variables by LSOA (cont.)

13. Plus other LSOA-level user-requested data commissioned from the ONS
    e.g. based on 2011 census questionnaire
    detail depends on how invasive or sensitive data are

IMD changes through time: 2015 to 2019

Table 2: Number of neighbourhoods in each decile of the IMD2019 and the IMD2015

Number of Lower-layer Super Output Areas		Index of Multiple Deprivation 2015										Total
		Most deprived 10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	Least deprived 10%	
Index of Multiple Deprivation 2019	Most deprived 10%	**2883**	400	1	0	0	0	0	0	0	0	*3284*
	10-20%	395	**2316**	567	6	0	0	0	0	0	0	*3284*
	20-30%	6	545	**2073**	643	18	0	0	0	0	0	*3285*
	30-40%	0	22	612	**1892**	726	31	1	0	0	0	*3284*
	40-50%	0	1	32	663	**1834**	721	31	3	0	0	*3285*
	50-60%	0	0	0	76	652	**1838**	685	33	0	0	*3284*
	60-70%	0	0	0	3	49	641	**1833**	719	38	1	*3284*
	70-80%	0	0	0	0	6	51	682	**1862**	671	13	*3285*
	80-90%	0	0	0	1	0	2	51	650	**2076**	504	*3284*
	Least deprived 10%	0	0	0	0	0	0	1	18	499	**2767**	*3285*
	Total	*3284*	*3284*	*3285*	*3284*	*3285*	*3284*	*3284*	*3285*	*3284*	*3285*	***32844***

Source: Office for National Statistics
The English Indices of Deprivation 2019 (IoD2019)

Actuarial Research Centre
Institute and Faculty of Actuaries

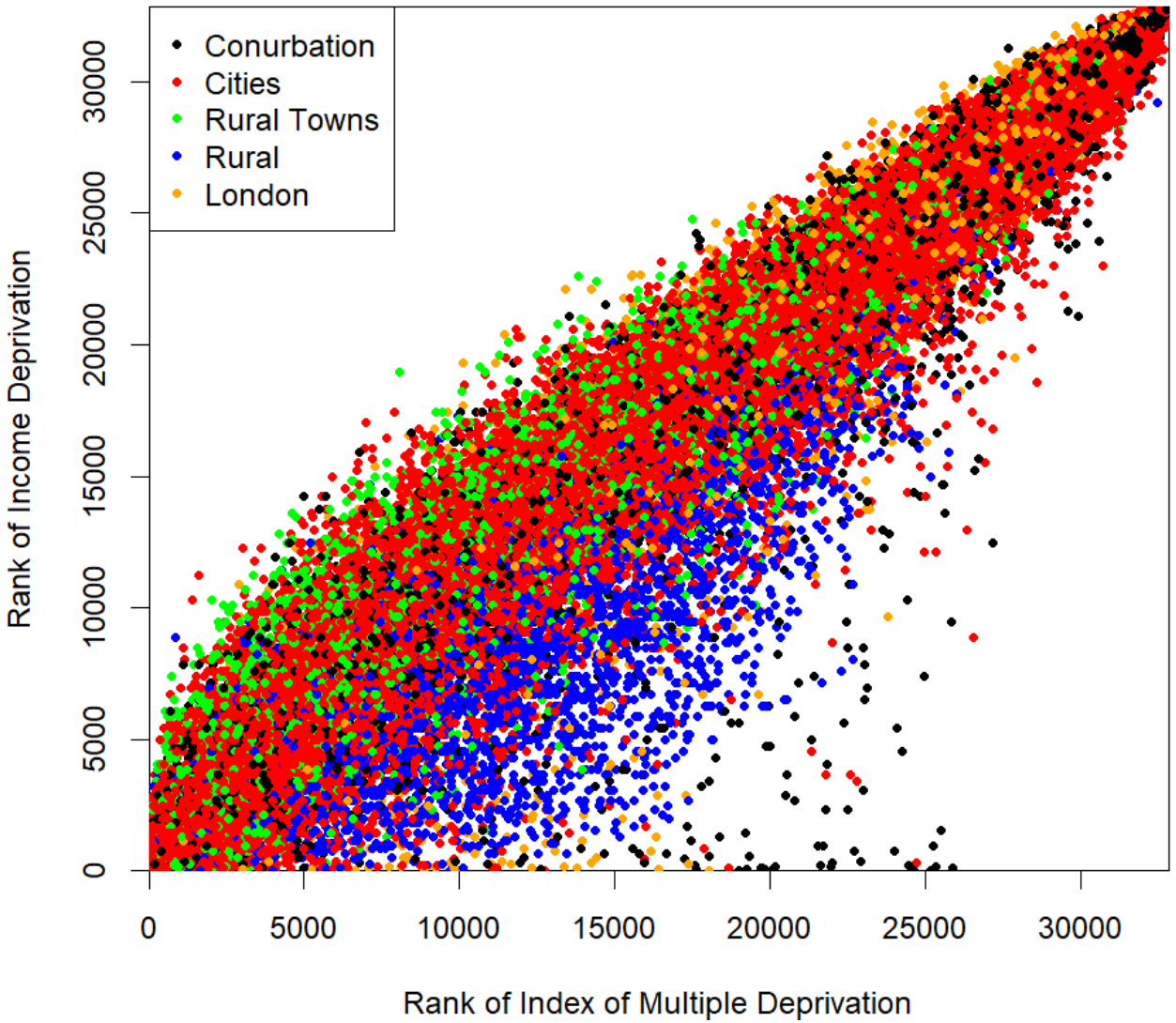
Scatterplots of pairs of predictive variables

- E.g. predictive variables (X_i, Y_i), for $i = 1, \ldots, L = 32844$
- R_{Xi} = rank of X_i out of $X_1, \ldots, X_L$
- R_{Yi} = rank of Y_i out of $Y_1, \ldots, Y_L$
- Scatterplot (R_{Xi}, R_{Yi}) for $i = 1, \ldots, L = 32844$
  $\Rightarrow$ focus on the dependency between X and Y
- When choosing which predictive variables to use, *avoid* pairs that are very highly correlated.

- Scatterplots can be coloured e.g. by urban rural group
  $\Rightarrow$ insights into what characterises different urban-rural classes

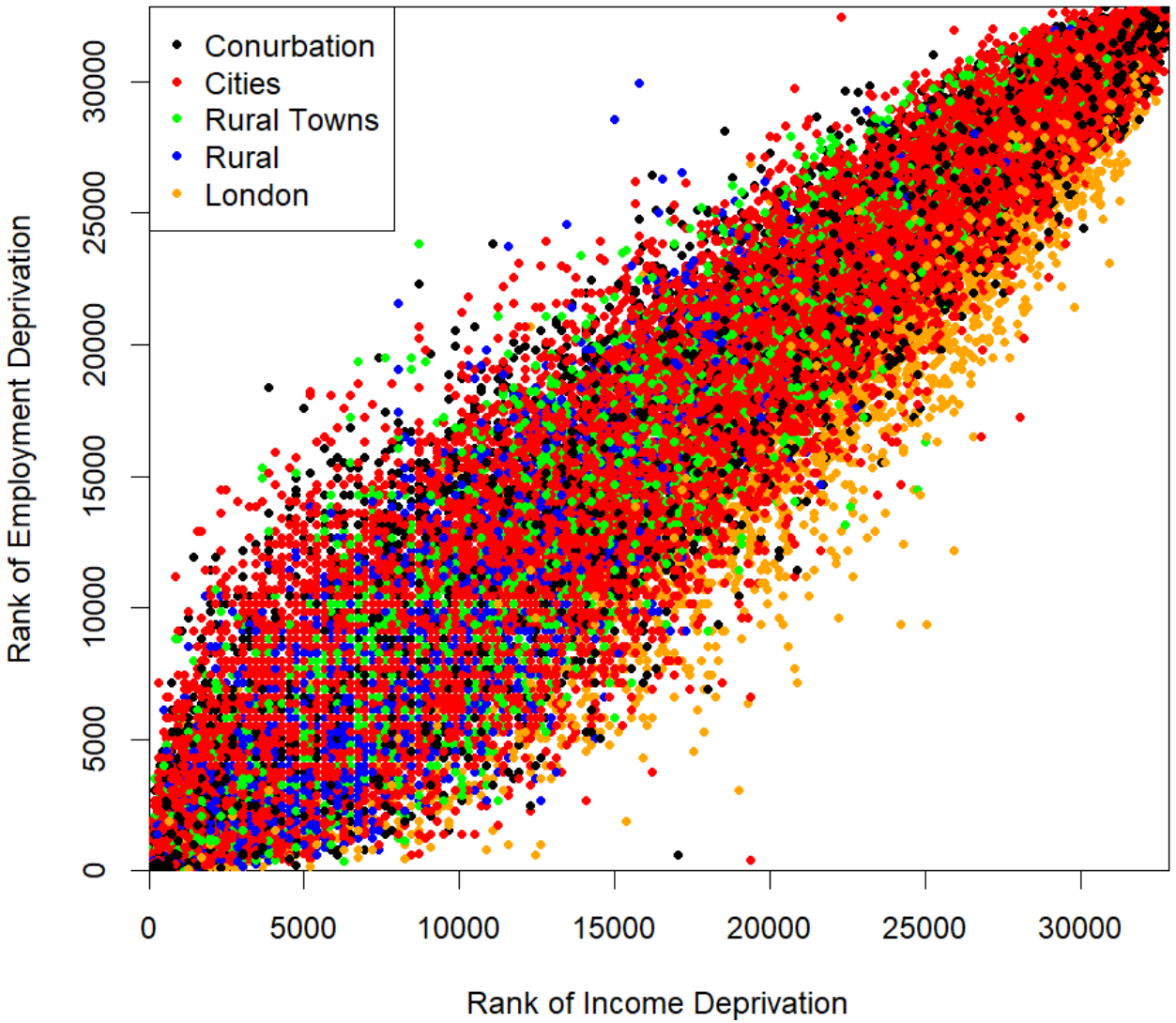
Actuarial Research Centre
Institute and Faculty of Actuaries

Index of Multiple Deprivation vs Income Deprivation



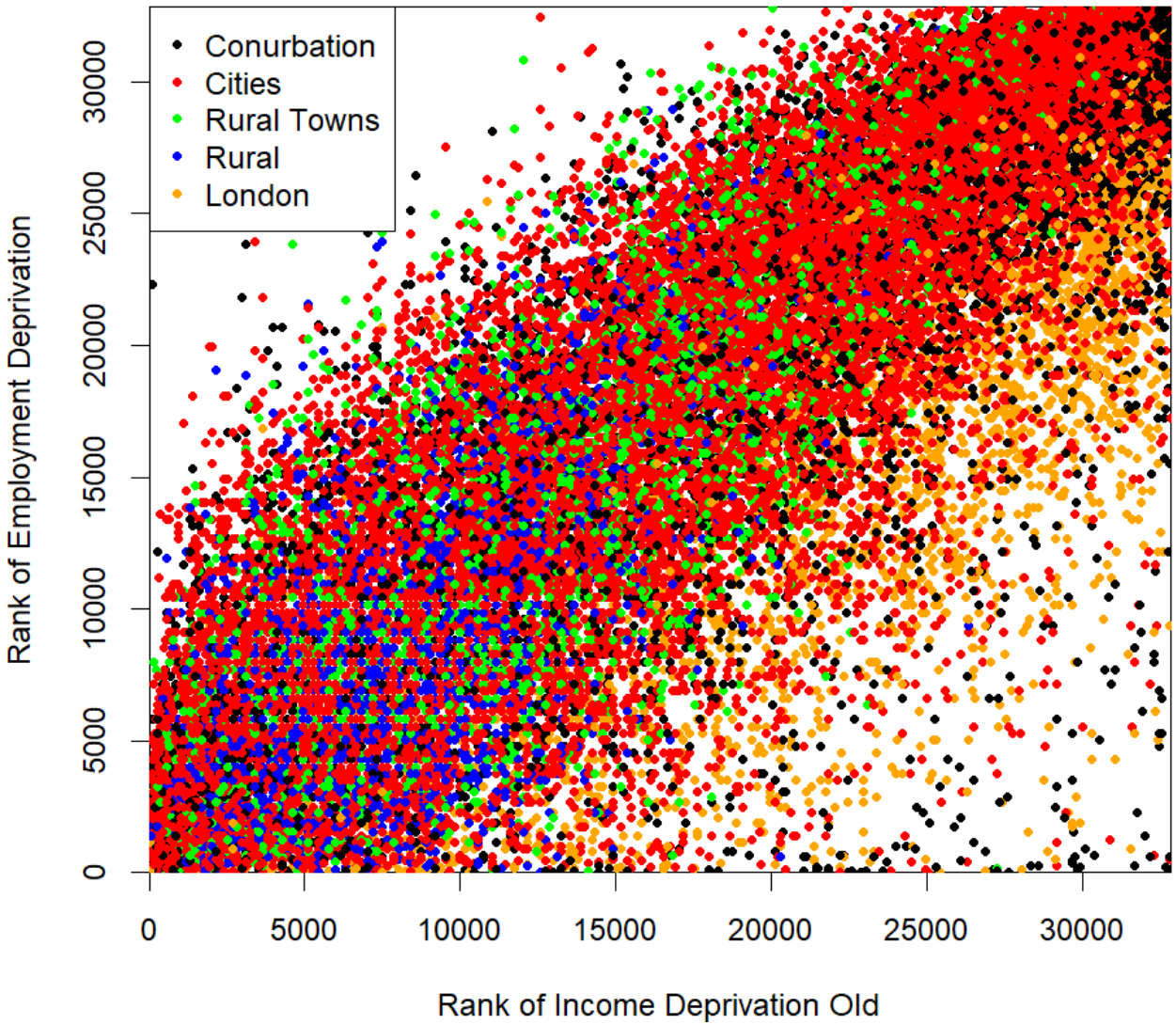
Actuarial Research Centre
Institute and Faculty of Actuaries

Income Deprivation vs Employment Deprivation



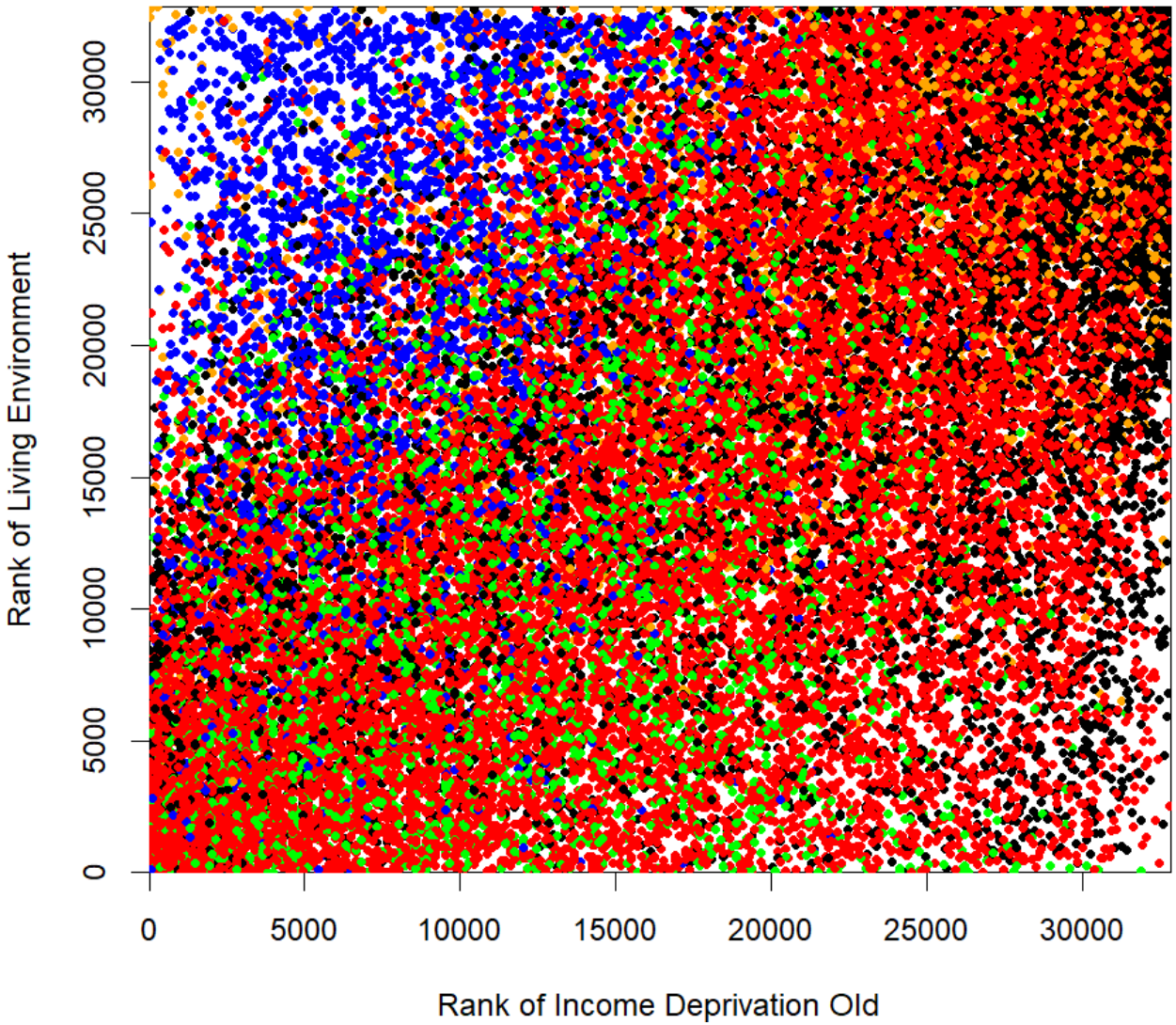
Actuarial Research Centre
Institute and Faculty of Actuaries

Income Deprivation Old vs Employment Deprivation



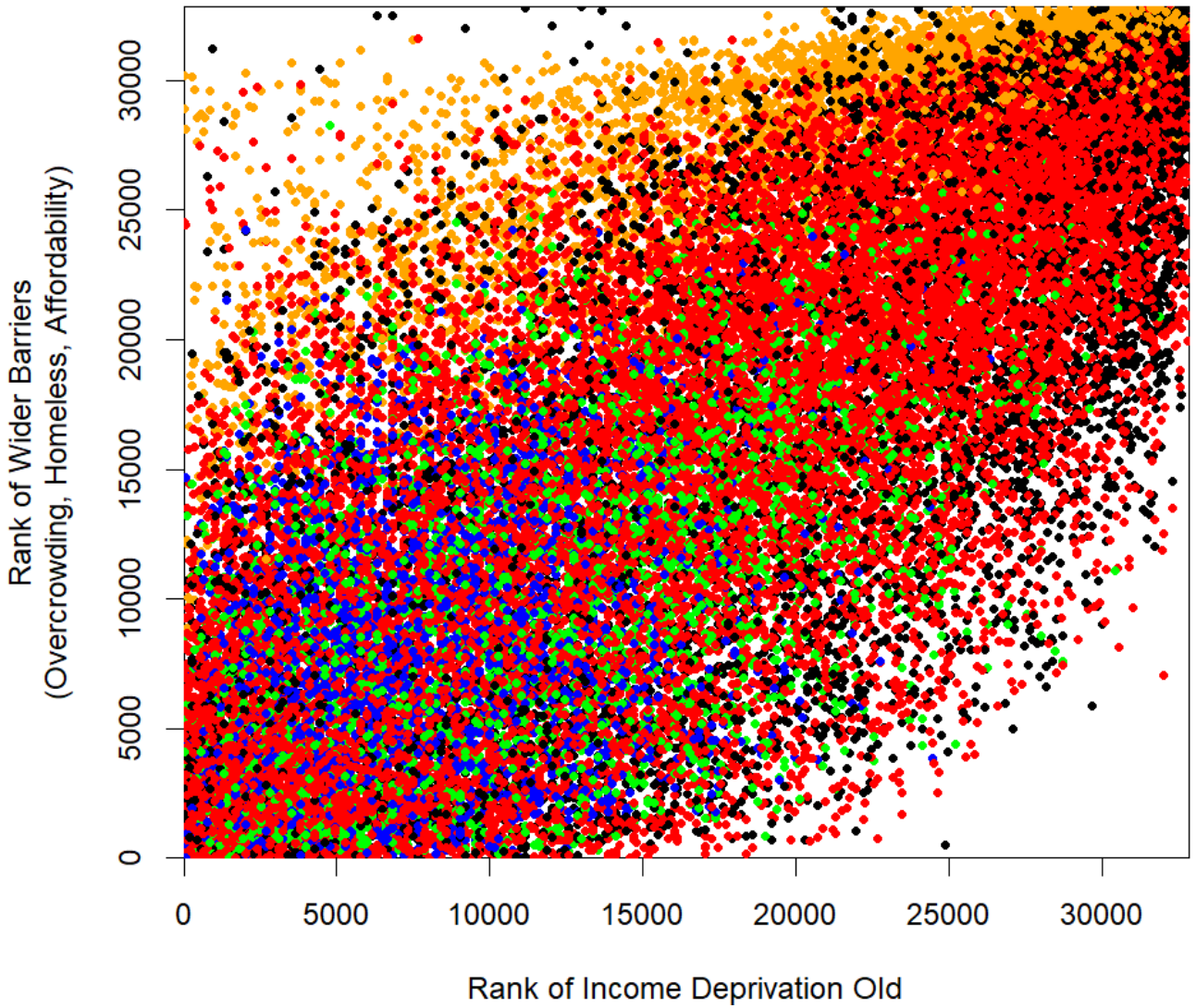
Actuarial Research Centre
Institute and Faculty of Actuaries

Income Deprivation Old vs Living Environment



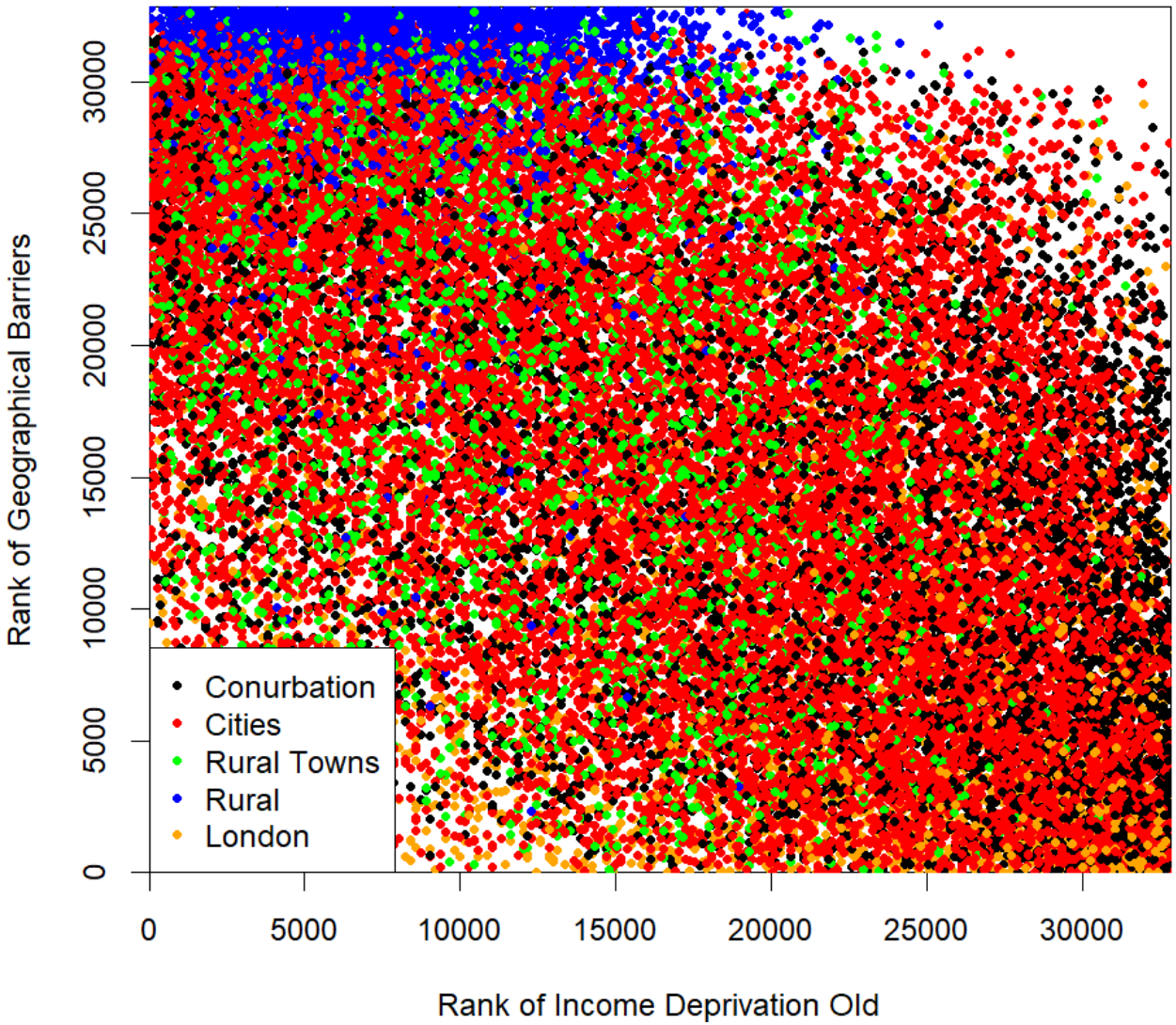
Actuarial Research Centre
Institute and Faculty of Actuaries

Income Deprivation Old vs Wider Barriers



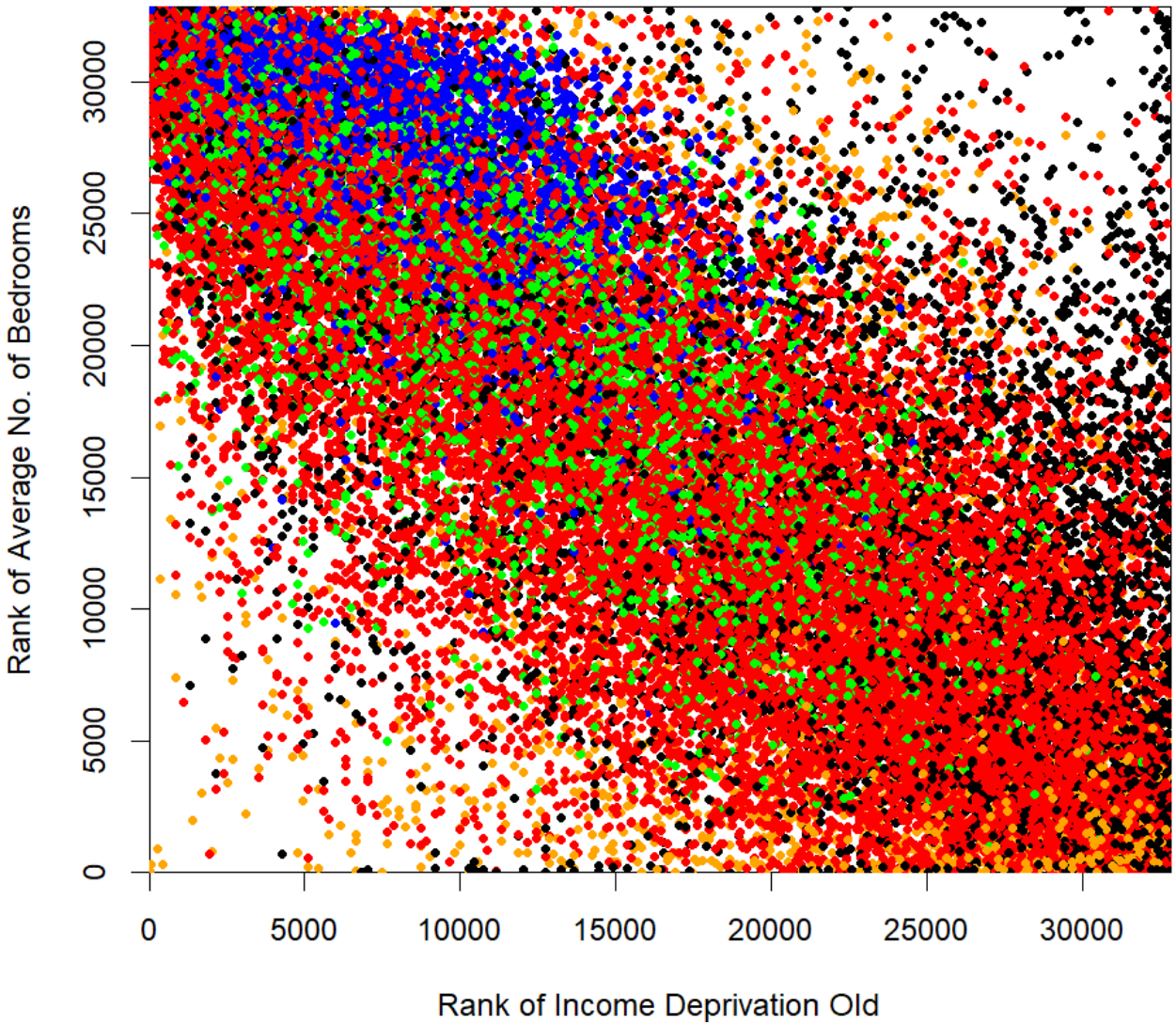
Actuarial Research Centre
Institute and Faculty of Actuaries

Income Deprivation Old vs Geographical Barriers



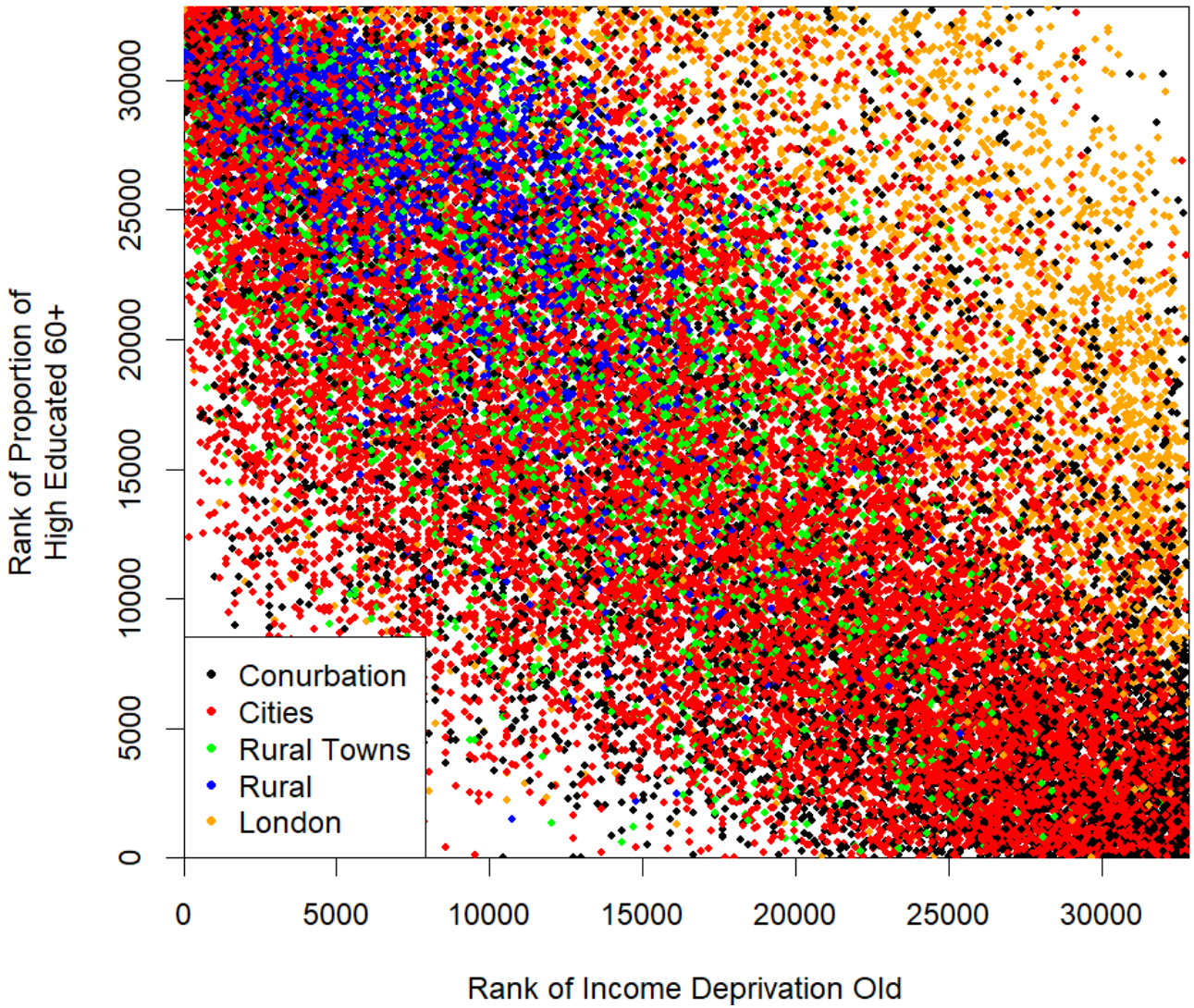
Actuarial Research Centre
Institute and Faculty of Actuaries

Income Deprivation Old vs Average Bedrooms



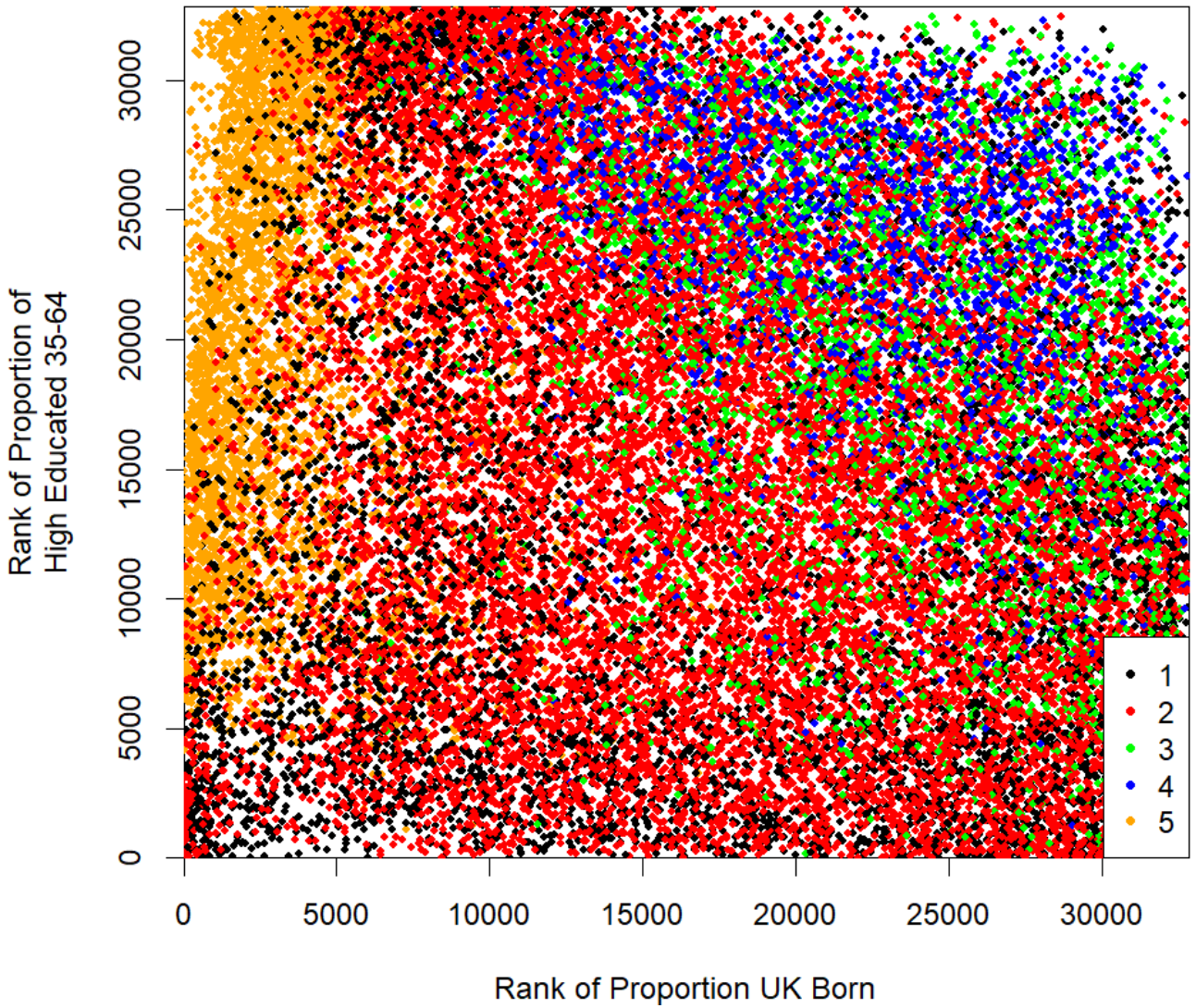
Actuarial Research Centre
Institute and Faculty of Actuaries

Income Deprivation Old vs High Educated 65+



Actuarial Research Centre
Institute and Faculty of Actuaries

UK Born vs High Educated 35-64



Actuarial Research Centre
Institute and Faculty of Actuaries

Cause of death data

- All-cause $\Rightarrow D(g, i, t, x)$ by LSOA
  $\Rightarrow$ lots of 0's and 1's
  Not considered to be invasive
- Cause of death:
  small numbers are considered to be invasive/sensitive
- Death counts: $D(g, r, i, c, t, x)$
  - g: gender (2)
  - r: region (9)
  - i: income deprivation decile (10)
  - c: cause of death (34)
  - t: year (16)
  - x: 5-year age groups

Actuarial Research Centre
Institute and Faculty of Actuaries

Questions to be addressed

- What are good stochastic mortality models for capturing differences between deprivation deciles?
- What are the most significant socio-economic factors that influence mortality rates?
- What other factors push mortality rates up or down at the level of small geographical or regional level?
- Does it make a difference if a neighbourhood is in an urban or rural area?
- After socio-economic and non-spatial effects have been filtered out, what remains in terms of spatial or regional variation in mortality across England.
- How much inequality is there in mortality rates at different ages?

Questions to be addressed (cont.)

- What is the difference between controllable and non-controllable risk factors?
- Which causes of death have significant controllable risk factors?
- Which causes of death have significant levels of mortality inequality?

- What are the contributors to the slowdown in mortality improvements since 2011?

Actuarial Research Centre
Institute and Faculty of Actuaries

Questions?

E: A.J.G.Cairns@hw.ac.uk

Find out more:

ARC website: www.actuaries.org.uk/ARC

Project website: www.macs.hw.ac.uk/∼andrewc/ARCresources

HERIOT WATT UNIVERSITY

SOCIETY OF ACTUARIES

Canadian Institute of Actuaries | Institut canadien des actuaires

Actuarial Research Centre
Institute and Faculty of Actuaries