Using Local Linear Regression to Model Socio-Economic and Geographical Effects

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

Outline

- Background
- Data – English mortality
- Methodology
- Results and discussion

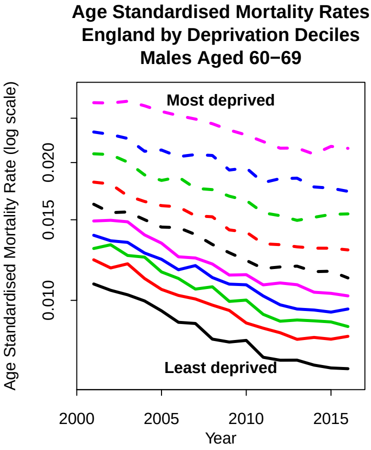
Actuarial Research Centre
Institute and Faculty of Actuaries

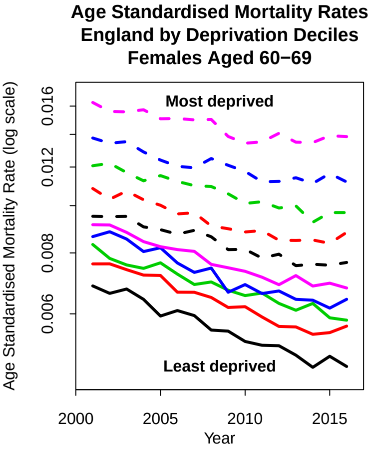
Background

- Considering here:
  male mortality in England
  (results for females similar and consistent)
- Stylised facts:
  - Mortality varies by socio-economic group
  - Mortality varies by region

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.

Data: LSOA's

- England only
- Lower Layer Super Output Areas: LSOA's
- $L = 32,844$ small geographical areas
- Socio-economically homogeneous
- Average size $\approx$ 1600 persons
- LSOA's $i = 1, \ldots, L$,
  single years ($t =$2001-2016), single ages, x:
  - Deaths: $D(i,t,x)$
  - Exposures: $E(i,t,x)$ (population)
- Plus many *static* predictive variables for each LSOA

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
- Urban/rural classification (categorical)
- ....

Methodology

- $D(i,t,x)$, $E(i,t,x)$ deaths and exposures by LSOA
- National death rates (all t and x)

$$m(t,x) = \frac{\sum_{i=1}^{L} D(i,t,x)}{\sum_{i=1}^{L} E(i,t,x)}$$

- LSOA's ($i = 1, \ldots, L$) local death rates: $m(i,t,x)$
  General Model: $\mathrm{E}[D(i,t,x)] = m(i,t,x)E(i,t,x)$
  How to model $m(i,t,x)$?

Methodology (cont.)

General approach:

- Over a limited age range (e.g. 60-69); and
- Over a (potentially) limited range of years:

$$m(i,t,x) = m(t,x)F_1(i)F_2(i)$$

- $F_1(i)$ = relative risk due to socio-economic characteristics
  - GLM
  - kernel smoothing
  - local linear regression
- $F_2(i)$ = additional relative risk capturing spatial effects
  - kernel smoothing

Methodology (cont.)

- Years: $t = t_0, \ldots, t_1$
- Ages: $x = x_0, \ldots, x_1$
- Actual deaths by LSOA

$$D(i) = \sum_{t=t_0}^{t_1} \sum_{x=x_0}^{x_1} D(i,t,x)$$

- Expected deaths by LSOA (no modelled effects)

$$\hat{D}_0(i) = \sum_{t=t_0}^{t_1} \sum_{x=x_0}^{x_1} m(t,x)E(i,t,x)$$

- Actual-over-expected by LSOA

$$R_0(i) = D(i)/\hat{D}_0(i)$$

Stage 1: Introduce Predictive Variables

- LSOA's: $i = 1, \ldots, L$
- Predictive variables (PV): $j = 1, \ldots, n_P$
- $P(i,j) =$ unadjusted PV
- Different PVs are on different scales
  (e.g. $[0,1]$, $[0,100]$, $(-\infty, +\infty)$)
- Hence: for each j standardise each PV ($i = 1, \ldots, L$)

$$P(i,j) \longrightarrow X(i,j) = \frac{P(i,j) - \bar{P}(i,j)}{S.D.P(i,j)}$$

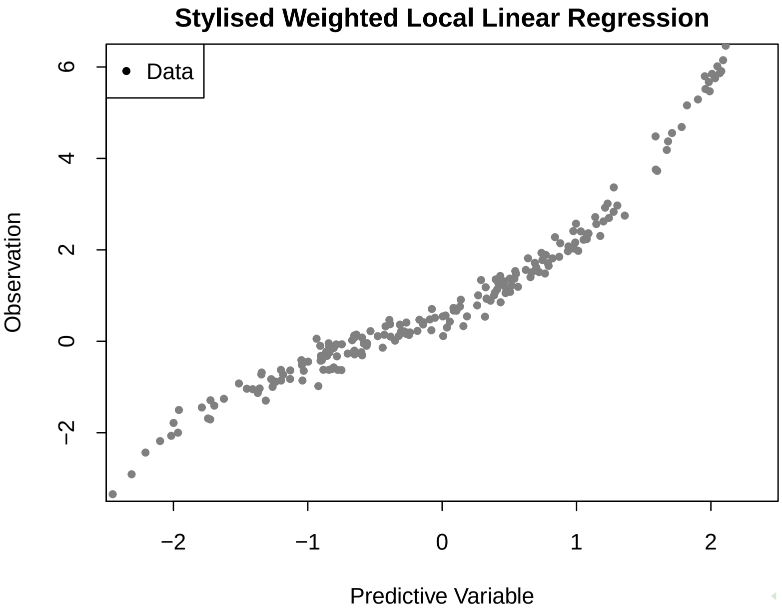
  So each $X(i,0$ has mean 0 and variance 1.

- Purpose of standardisation:
  Simplifies the system of weighting later in Stage 1
- Vector: $X(i) = (X(i,1), \ldots, X(i,n_P))'$

Stage 1: Urban versus Rural

- Urban-rural classification
  - 1: Conurbation: not London (7921)
  - 2: City or town (14515)
  - 3: Rural town (3056)
  - 4: Rural village and dispersed (2542)
  - 5: Conurbation; London (4810 LSOA's)
- Preliminary experiments $\Rightarrow$
  contribution and importance of specific predictive variables varies significantly between urban and rural LSOA's
- Hence: incorporate urban/rural classification into the process.

Stage 1: Local Linear Regression



Stage 1: Local Linear Regression

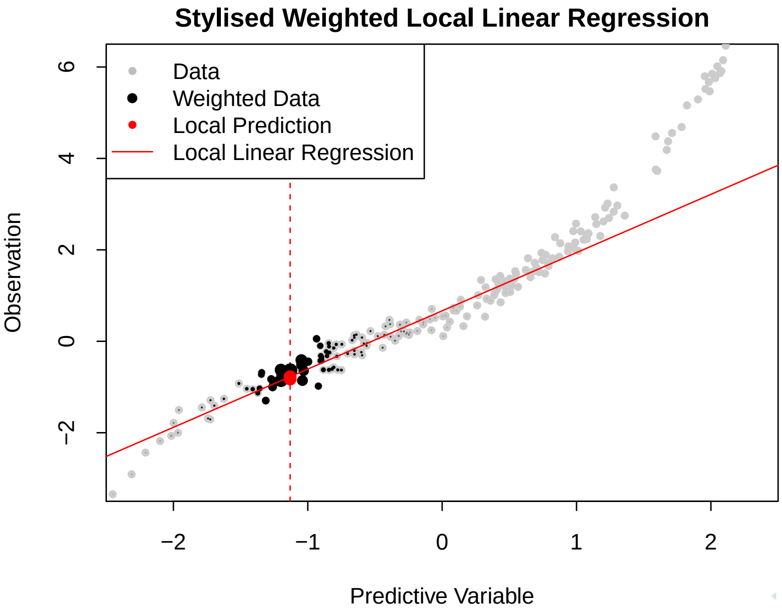
Stylised Example: X one dimensional

- Observe $(X(i), Y(i)), \quad i = 1, \ldots, n$
- What is $\hat{Y}(i) = E[Y(i)|X(i)]$?
- Weighted least squares:

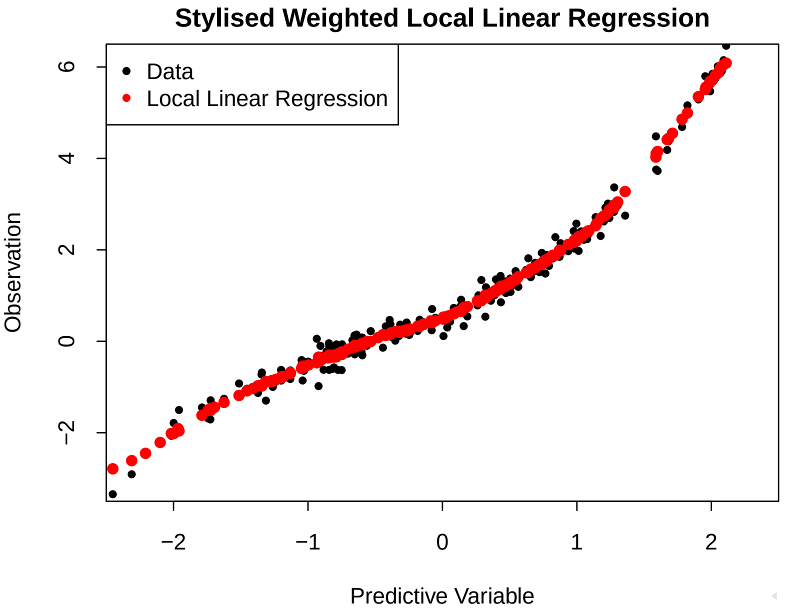
$$\text{minimise} \quad S_i = \sum_{j=1}^{n} w(i,j)\,(Y(j) - (a + bX(j)))^2$$

- Weights, $w(i,j) \to 0$ as $X(j)$ gets further from $X(i)$
  $\Rightarrow$ fit a straight line through points near $X(i)$
- Minimisation $\Rightarrow \hat{a}(i), \hat{b}(i)$
- $\hat{Y}(i) = \hat{a}(i) + \hat{b}(i)X(i)$
- Could also use e.g. B-splines
  But might not be practical if X has several dimensions.

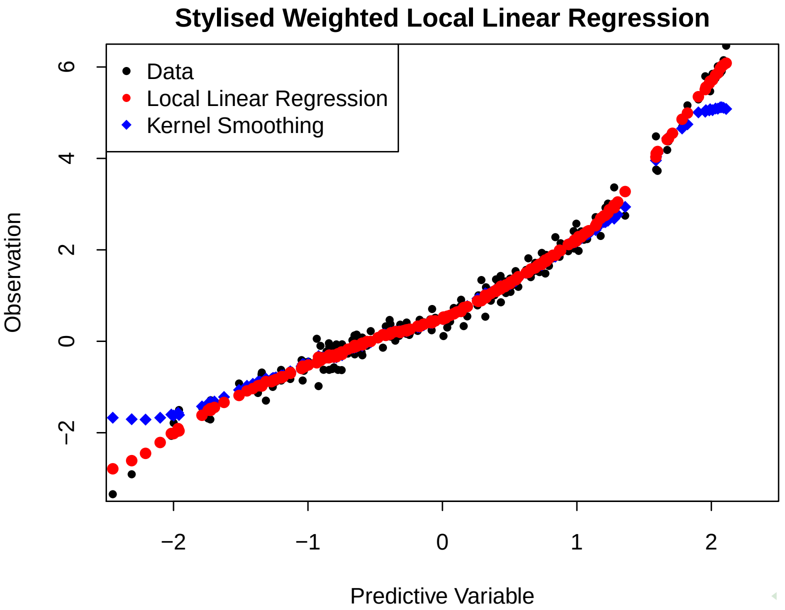
Stage 1: Local Linear Regression



Local Linear Regression



Stage 1: Local Linear Regression



Stage 1: Local Linear Regression

- For each LSOA, i
- Estimate the socio-economic-specific Relative Risk, $F_1(i)$
- For each i, fit an n_P-dimensional sheet around $X(i)$

$$F(i, \mathbf{x}) = a(i) + \mathbf{b}(i)^T \mathbf{x}$$

- n_P predictive variables exclude urban-rural classification urban-rural handled in the weights, $w_1(i, j)$
- Minimise

$$S(a(i), b(i)) = \sum_j w_1(i, j) \left(R_0(j) - a(i) - b(i)^T X(j) \right)^2$$

over $a(i)$ and $b(i)$

Stage 1: Local Linear Regression (cont.)

- Then set

$$F_1(i) \quad = \quad a(i) + b(i)^T X(i)$$

$\Rightarrow$ relative risk accounting for socio-economic factors

- Update estimated deaths:

$$\hat{D}_1(i) \quad = \quad \hat{D}_0(i) F_1(i)$$

Stage 1: Local Linear Regression (cont.)

How to calculate the weights?

- $w(i,i) = 0$
- $w(i,j) = 0$ if LSOA's i and j are in different urban-rural groups

  Otherwise:
- $w(i,j)$ depends on the "distance" between predictive variables $X(i)$ and $X(j)$
- $w(i,j) \to 0$ as the distance gets larger
- We also give greater weights to LSOAs that have higher expected deaths, $\hat{D}_0(j)$
  $\Rightarrow$ more reliable A/E, $R_0(j)$

Stage 1 $\rightarrow$ Stage 2

$$
\begin{aligned}
D(i) &= \text{LSOA actual deaths} \\
\hat{D}_0(i) &= \text{LSOA expected deaths with no predictive variables} \\
\hat{D}_1(i) &= \text{LSOA expected deaths with predictive variables} \\
R_1(i) &= \frac{D(i)}{\hat{D}_1(i)} = \text{updated actual-over-expected}
\end{aligned}
$$

Stage 2: Add location data:

$$
\begin{aligned}
Y(i) &= \text{LSOA location co-ordinates} \\
&= \text{(latitude, longitude)}
\end{aligned}
$$

Kernel smooth the $R_1(i)$ using location data.

Stage 2: Smooth A/E by Location

Simpler: use kernel smoothing

Estimate the *additional* location-specific relative risk

$$F_2(i) \;=\; \frac{\sum_j w_2(i,j) R_1(i)}{\sum_j w_2(i,j)}$$

Then the fitted expected deaths are

$$\hat{D}_2(i) = \hat{D}_0(i) F_1(i) F_2(i)$$

Weights, $w_2(i,j)$, depend on the physical distance between the two LSOA's

Data and Results

- 2001-2015
- Ages: 40-49, 50-59, 60-69, 70-79, 80-89
- Predictive variables:
  - income deprivation (elderly; receiving government benefits)
  - employment deprivation (unemployment)
  - average number of bedrooms
  - living environment deprivation (housing quality and air quality)
  - wider barriers (overcrowding)
  - high education (level 4+) amongst over 65's

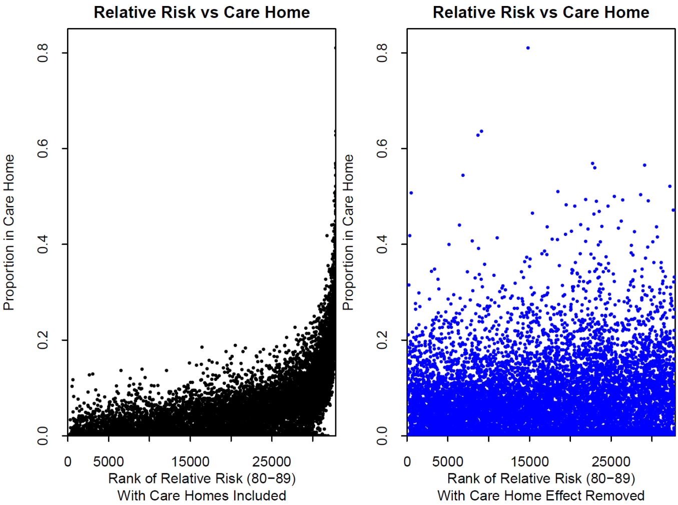
  - % in care home (60+ with nursing)
  - % in care home (60+ without nursing)

  - urban-rural classification

Role of Predictive Variables

- Employment deprivation is the main driver for younger age groups
- Income deprivation (elderly) is the main driver for older age groups
- Urban-rural classification is also an important driver
- Bedrooms, living environment, wider barriers and high education are second order but significant
- Care homes:
  - "nuisance" variables when considering socio-economic effects
  - but including these predictive variables is very important
  - methodology allows us to filter out the impact of care homes on individual LSOA mortality
  - E.g. males 80-89 in a care home with nursing:
    mortality is 3x to 6x higher than not in a care home

Where are the care homes?

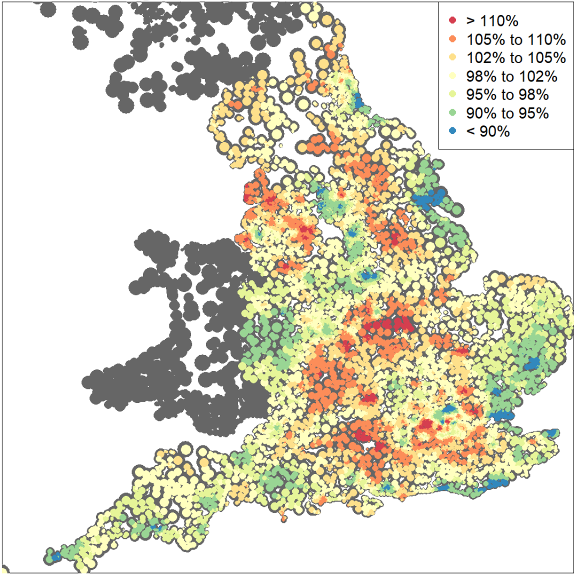


Location-Specific Relative Risk

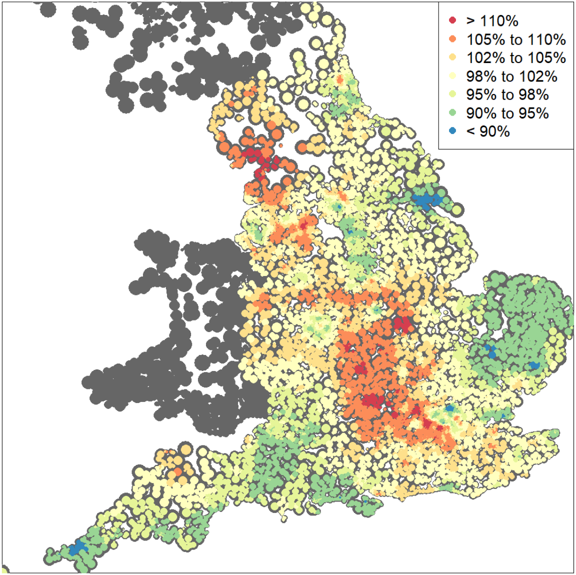
$$F_2(i) = \frac{\sum_j w_2(i,j) R_1(i)}{\sum_j w_2(i,j)}$$

the residual risk after fitting socio-economic effects, $F_1(i)$

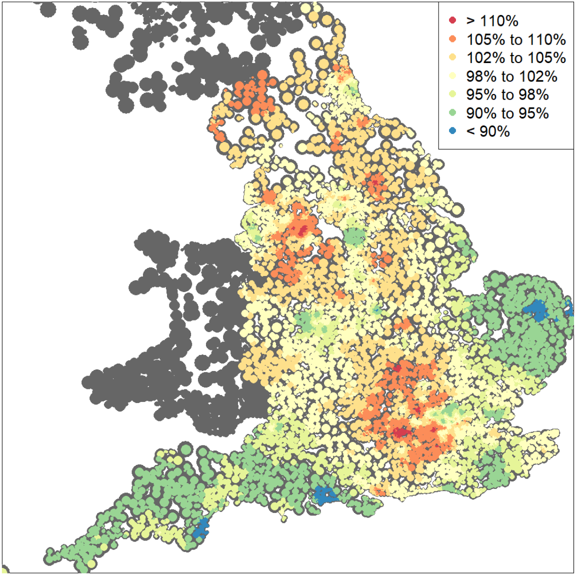
Location-Specific Relative Risk
England, Males 40-49, 2001-2015



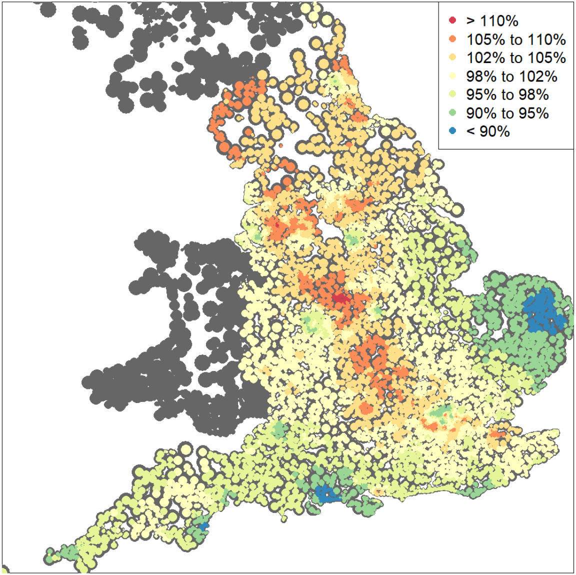
Location-Specific Relative Risk
England, Males 50-59, 2001-2015



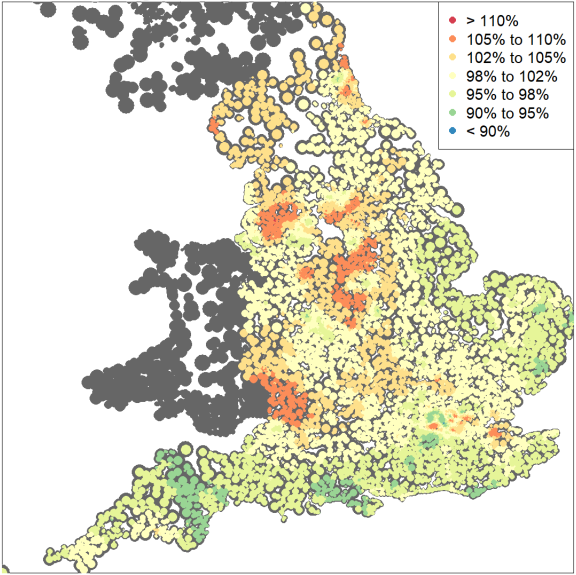
Location-Specific Relative Risk
England, Males 60-69, 2001-2015



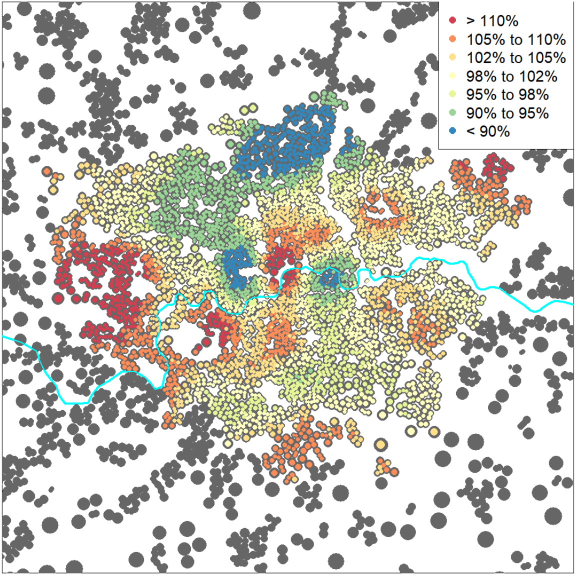
Location-Specific Relative Risk England, Males 70-79, 2001-2015



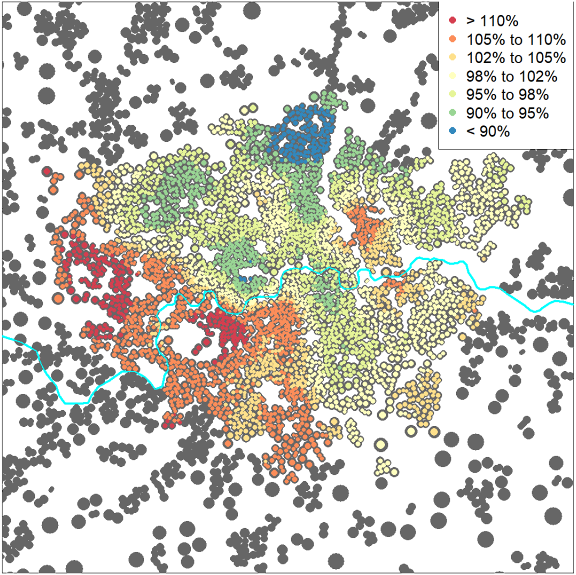
Location-Specific Relative Risk
England, Males 80-89, 2001-2015



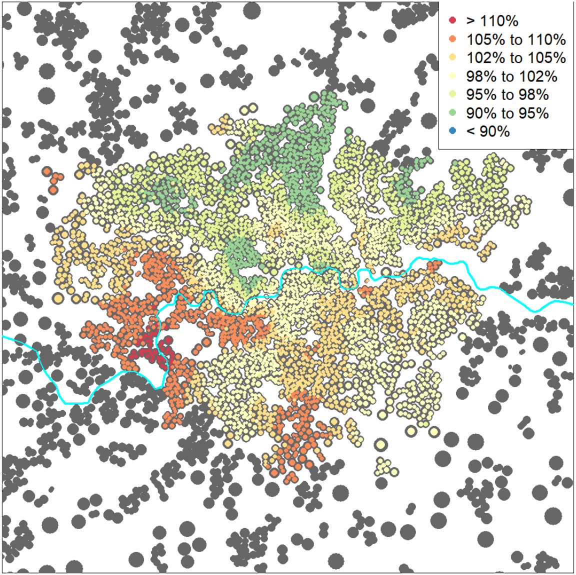
Location-Specific Relative Risk
London, Males 40-49, 2001-2015



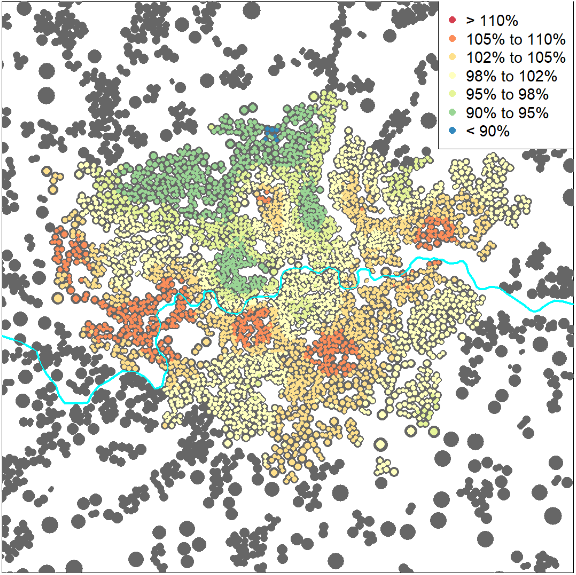
Location-Specific Relative Risk
London, Males 50-59, 2001-2015



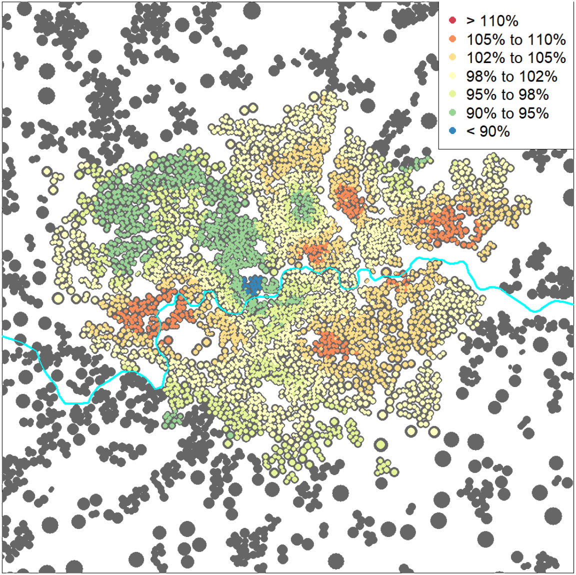
Location-Specific Relative Risk
London, Males 60-69, 2001-2015



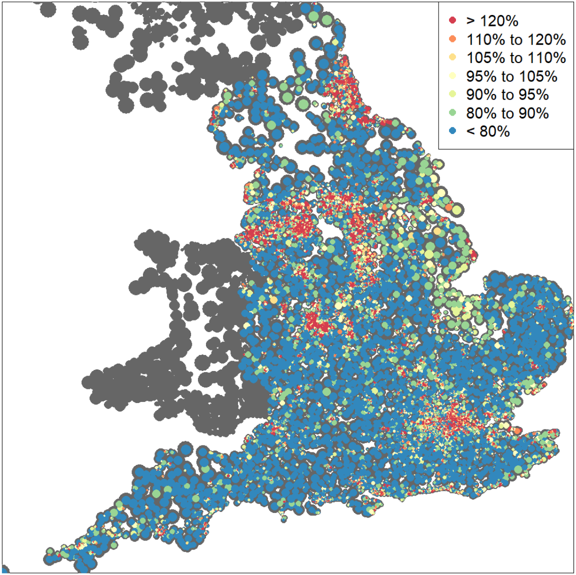
Location-Specific Relative Risk London, Males 70-79, 2001-2015



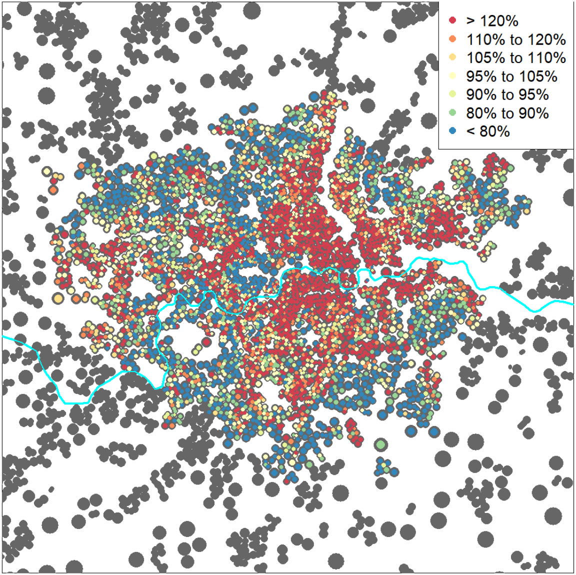
Location-Specific Relative Risk
London, Males 80-89, 2001-2015



Combined Relative Risk
England, Males 60-69, 2001-2015



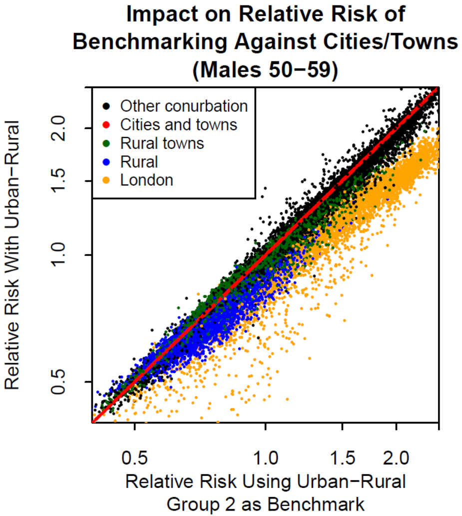
Combined Relative Risk
London, Males 60-69, 2001-2015

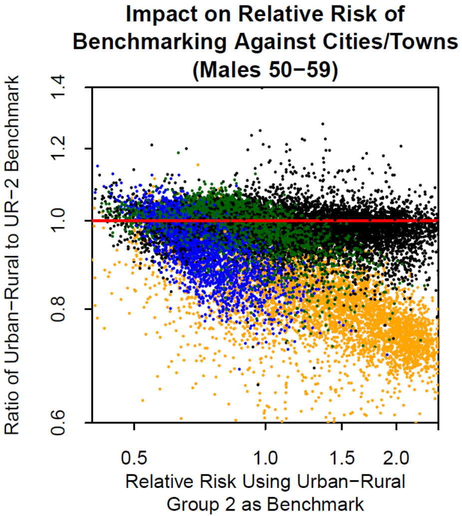


Benchmarking against UR-2 cities and large towns

- Previously: $w_1(i,j) > 0$ only if i and j in the same urban-rural class
- Experiment:
  Benchmark all LSOA's against the socio-economically nearest in urban-rural class 2 (cities and large towns)
  (Class 2 is the largest, and has the widest spread of socio-economic predictive variables)
  $w_1(i,j) > 0$ only if $u(j) = 2$
- Original: $\hat{F}_1(i)$
- Experiment $\Rightarrow \hat{F}_1^{UR2}(i)$
- Plot A: $\hat{F}_1^{UR2}(i)$ versus $\hat{F}_1(i)$
- Plot B: $\hat{F}_1^{UR2}(i)$ versus the Ratio $\hat{F}_1(i)/\hat{F}_1^{UR2}(i)$

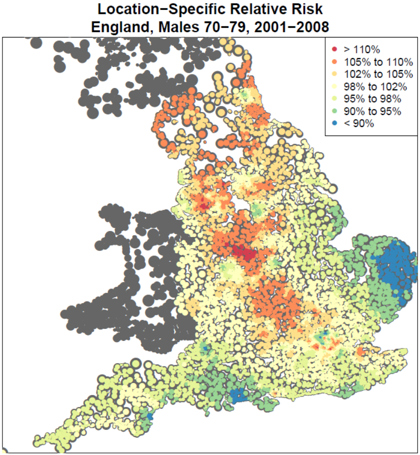
Benchmarking against UR-2 cities and large towns

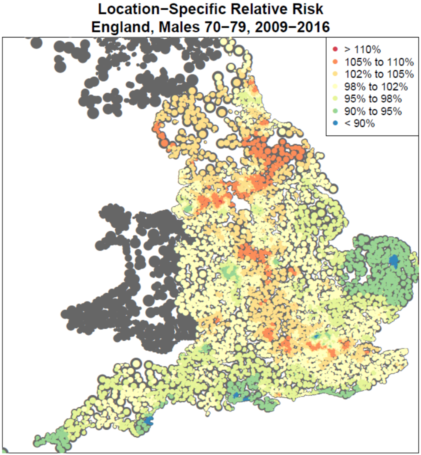




Rural (blue) and London (orange):
on a like-for-like basis, much lower mortality than cities and large towns

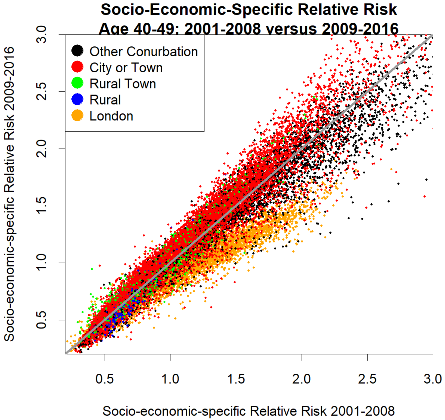
2001-2008 versus 2009-2016

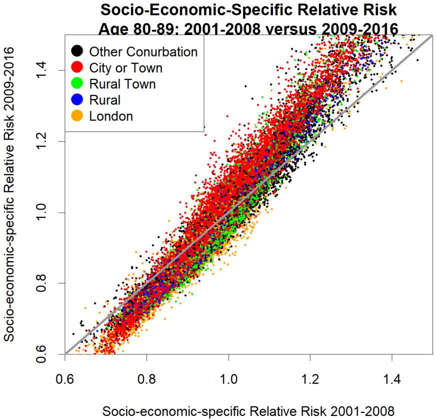




Some variation over time.
Location is becoming less important over time.

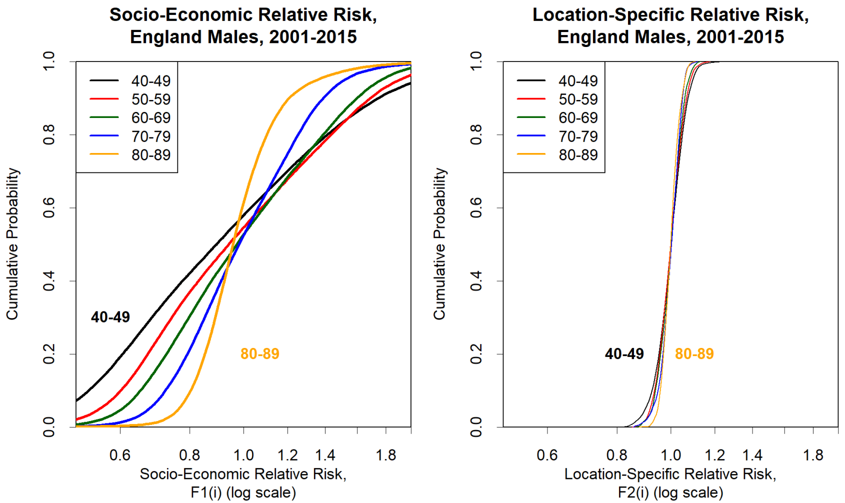
2001-2008 versus 2009-2016: Ages 40-49 and 80-89





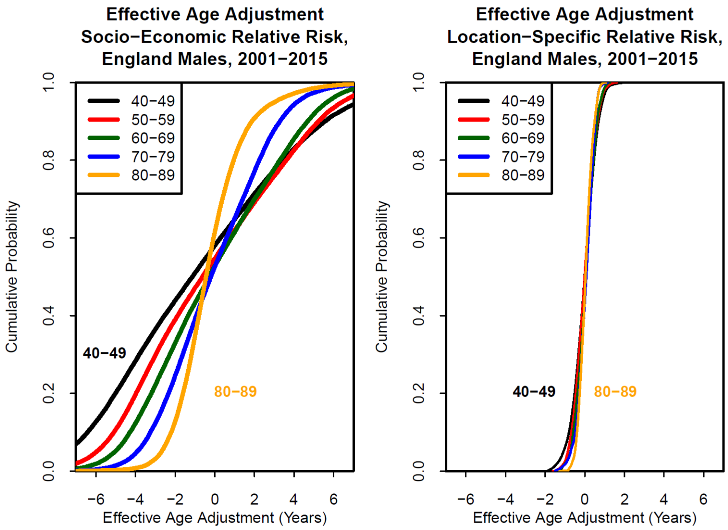
- Sampling variation is significant
- Widening inequality gap at 80-89
- Stable gap at 40-49, except London: narrowing gap

Socio-Economic vs Spatial Effects



- Location contributes 1.3% to 3.5% of the variance in the relative risk

Socio-Economic vs Spatial Effects



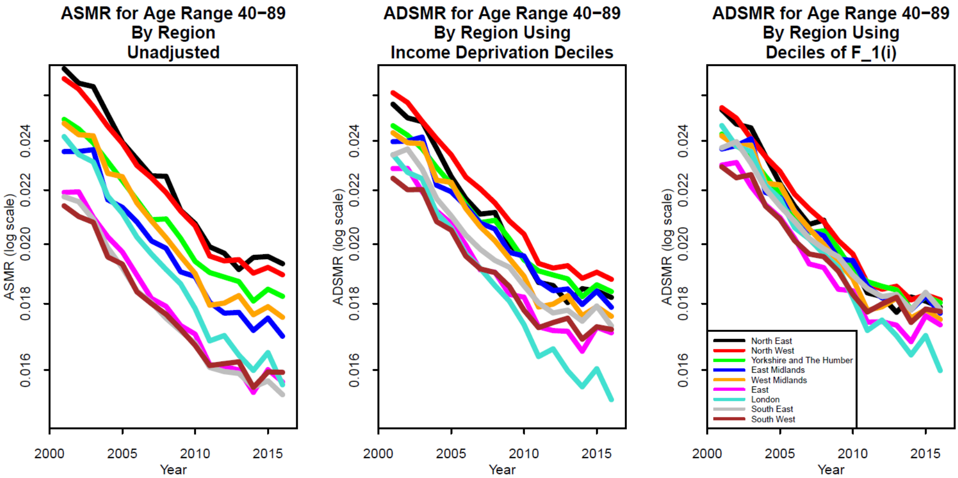
e.g. Effective age adjustment$= -4 \Rightarrow$ mortality is as if 4 years younger

Actual-over-expected: Ages 60-69

Region	No effect	Socio-economic only
North East	118	100
North West	116	102
Yorkshire and The Humber	107	100
East Midlands	98	100
West Midlands	105	99
East	88	96
London	105	100
South East	89	101
South West	87	94

- Similar patterns for other age groups and for females

Use $\hat{F}_1(i)$ to create deciles



Regional differences narrow, but more obvious London effect

Significant improvement over income deprivation deciles

Conclusions

- Key predictive variables: income and employment deprivation
- But other predictive variables play important roles
- Socio-economic relative risk, $F_1(i)$, outperforms *income deprivation* as a predictor

- Spatial/regional effects are significant
- But much less important than socio-economic (non-regional) effects
- Next steps:
  - Both effects: can these be used to improve predictions of insurance and pensions mortality?
  - More detail:
    Sessional research meeting on 6 January 2020

HERIOT WATT UNIVERSITY

Thank You!

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