HERIOT WATT UNIVERSITY

Actuarial Research Centre
Institute and Faculty of Actuaries

Drivers of Mortality: Risk Factors and Inequality

Andrew J.G. Cairns[a], [b]Torsten Kleinow[c]and Jie Wen[d]

[a]Maxwell Institute for Mathematical Sciences, and Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, UK.
[b]Corresponding author: A.J.G.Cairns@hw.ac.uk
[c]Research Centre for Longevity Risk, Faculty of Economics and Business, University of Amsterdam
[d]Lloyds Banking Group

Drivers of Mortality: Risk Factors and Inequality

Andrew J.G. Cairns[a], [b]Torsten Kleinow[c]and Jie Wen[d]

This version: March 31, 2023

Abstract

This paper takes a detailed look at socio-economic variation in mortality across England. Local linear regression is used to analyse all-cause mortality at neighbourhood level (Lower Layer Super Output Areas) with mortality rates linked to a number of socio-economic predictive variables that determine the character of a neighbourhood. We find that old-age-income and employment deprivation are key determinants of mortality, but also that urban-rural class and the presence of care homes in a neighbourhood have an important role to play in assessing underlying mortality rates relative to national mortality. Remaining spatial/regional variation in mortality is found to be much less important than socio-economic variation and much lower than the remaining regional variation that results from the commonly-used Index of Multiple Deprivation.

Keywords: Mortality inequality; Lower Layer Super Output Area; Age and Deprivation Standardised Mortality Rate; Regional mortality variation; Local linear regression.

1 Introduction

1.1 Socio-economic differences in mortality

It is well known that socio-economic status and associated variables are strongly correlated with high and low mortality. The evidence base for this has been growing over a number of decades as data become available with sufficient detail to be able to investigate the dependency between particular predictive variables and mortality. Socio-economic status is not always easily available in the same form but typical information about individuals or groups includes income, income deprivation or affluence (e.g. Chetty et al., 2016 [US], Villegas and Haberman, 2014, Longevity Science Panel, 2018 [UK], Cairns et al., 2019 [Denmark], Wen et al., 2020, 2021 [UK, Canada]) and education (e.g. Mackenbach et al., 2003, 2016).

This paper seeks to delve more deeply into the links between different measures of socio-economic status, geographic location and mortality. We do this by exploiting a large dataset for England built up from multiple datasets obtained from the UK's Office for National Statistics (ONS).

A key modelling tool for this study is *local linear regression*. As a statistical method, a version was first proposed by Cleveland (1979) and is often referred to as LOWESS (Locally Weighted Scatterplot Smoothing). As we demonstrate in this paper, it is an effective

[a]Maxwell Institute for Mathematical Sciences, and Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, UK.

[b]Corresponding author: A.J.G.Cairns@hw.ac.uk

[c]Research Centre for Longevity Risk, Faculty of Economics and Business, University of Amsterdam

[d]Lloyds Banking Group

tool to handle large datasets with many covariates for mortality modelling. Compared to e.g. Generalised Linear Models (see, for example, Macdonald et al., 2018) the model fits the data effectively without having to specify in advance the functional relationship between predictive variables (including interactions) and the outcome.

1.2 Questions addressed in this paper

This paper contains a wide ranging analysis of all-cause mortality data addressing the following questions:

- What are the most significant socio-economic factors that influence mortality rates in England, and which predictive variables are the best at predicting high or low mortality at specific ages?
- How many predictive variables do we need to get a reasonable model for the mortality in small homogeneous neighbourhoods?
- Everything else being equal, does it make a difference if a neighbourhood is in an urban or rural area?
- Where are care homes located and what impact do they have on mortality?
- Can regional differences in mortality be explained entirely by differences in the socio-economic mix and other non-spatial predictive variables?
- 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?
- Have mortality inequalities been widening from 2001 to 2018?

In answering some of these questions, we propose a new method to estimate the relative risk that models the difference between neighbourhood and national mortality: with variation in the relative risk by age group and sex. The methodology can be adopted as a continuous index that actuaries and other stakeholders can use as a predictive variable alongside other variables that are available at the individual level such as pension level and geodemographic profiling (e.g. Richards, 2008). Additionally, it can be used to divide the population up into deciles instead of using the Index of Multiple Deprivation or income deprivation.

1.3 Outline of the paper

In Section 2 we introduce the several datasets that we will use in our analysis, with further details in Supplementary Appendix A. Section 3 sets the scene for what is to come with some introductory analysis of the data. This also introduces the problem of regional variation in mortality. Section 4 outlines the local linear regression methodology. Results and insights into mortality at the small-neighbourhood level are presented and discussed in Section 5. Section 6 concludes.

2 Data

Data used in this study have been sourced from the UK's Office for National Statistics (ONS), and are available at the level of Lower Layer Super Output Areas (LSOAs; small, socially-homogeneous, geographical areas with, on average, 1600 people) and are of the following types:

- mid-year population data by LSOA;
- all-cause death counts by LSOA;
- predictive variables by LSOA.

There are 32,844 LSOAs across England.

We outline key elements of the data below, with further details to be found in Supplementary Appendix A.

2.1 Population data

For each LSOA, we have mid-year population estimates from the ONS by single year of age and single calendar years from 2001-2018, $E(g, i, t, a)$, where g is the sex, i is the LSOA index, t is the calendar year, and a is the age last birthday. As with standard practice, we equate this with the central exposed to risk for age a last birthday across the whole of year t.

Exposures can be aggregated into specified deciles, regions or the whole of England as required.

2.2 Mortality data

For all cause mortality, the ONS make available death counts, $D(g, i, t, a)$, by sex g, LSOA i, single age a and single year t.

Death rates at the LSOA level are then

$$m_L(g, i, t, a) = \frac{D(g, i, t, a)}{E(g, i, t, a)}$$

and at the regional level for region r

$$m_R(g, r, t, a) = \frac{\sum_{i \in R(r)} D(g, i, t, a)}{\sum_{i \in R(r)} E(g, i, t, a)}$$

where the set $R(r)$ contains all LSOA's, i, that lie within region r.

Similar expressions can be written down for decile-based death rates, $m_D(g, j, t, a)$, for decile j, and national mortality, $m(g, t, a)$.

In the remainder of the paper we will assume that the mortality of men and women are modelled seperately. We will, therefore, drop g from our notation.

2.3 Migration

It needs to be noted that the exposure data do not contain information about migration between LSOAs or between regions[1]. This means that caution needs to be exercised when viewing death rates across age groups as there might be some variation in the underlying population mix.

2.4 Predictive variables

The focus of this paper is on identifying socio-economic and other variables that have significant predictive power in estimating mortality rates. A significant number of variables are available at the LSOA level. A key source of predictive variables is the 2015 edition of the Index of Multiple Deprivation (see ONS, 2015) and its domains and subdomains including income deprivation, employment deprivation, education deprivation, crime, barriers to housing and services, and living environment. Wen et al. (2021) investigated the impact of using IME-2004 as an alternative to IMD-2015. While, there were small differences in death rates between deciles their broad conclusions were not altered by the change. Additionally, we considered: average number of bedrooms; level of educational attainment by age group; occupation group by age group; proportion born in the UK; proportions, by age group, of people living in various types of communal establishment; region; and urban-rural class (UR1 conurbation non-London, UR2 city or town, UR3 rural town, UR4 rural village and dispersed, and UR5 conurbation London). Further details of these can be found in Supplementary Appendix A. The predictive variables are not time dependent, consistent with the model developed in Section 4.1.

3 Preliminary analysis

Before we move into a more detailed analysis on the various datasets, it is worthwhile summarising the main patterns that exist in socio-economic data. Building on the recommendations of the Longevity Science Panel (LSP 2018) we will, here, use income deprivation rather than the main IMD to divide the LSOAs into deciles. Income deprivation is, of course, not a causal effect, but it was found by LSP to be more strongly *correlated* with high and low mortality than other domains of the IMD.[2]

In the left-hand panel of Figure 1 we have plotted age-specific death rates in 2018 for each of the income-deprivation deciles. For both males and females (not plotted) we can make the following observations:[3]

- Even though the data contain a certain amount of sampling variation, the deciles are clearly ranked from decile 1 (most deprived, highest mortality) down to decile 10 (least deprived, lowest mortality).

[1]However, it is possible to assess net migration by comparing changes in exposures by cohort with death counts but not determine where individuals are coming from or going to.

[2]Note that the IMD itself includes a health domain. As with LSP (2018) we do not consider the health domain. Instead, we seek to identify socio-economic covariates that are predictors of poor health and increased mortality.

[3]For a more detailed analysis of English deciles, see Wen et al. (2021).

- There are very significant differences between the deciles at ages 40 to 60 (the mortality inequality gap) before gradually narrowing as the population gets older. This narrowing is very typical for mortality differences between socio-economic groups using different measures and in different countries (see, for example, Cairns et al., 2019, Wen et al., 2020, Redondo Lourés and Cairns, 2020).

  At ages 40 to 50, death rates in the most deprived group are around 4 times (males) and 3 times (females) the corresponding death rates in the least deprived group. Both narrow to about 1.4 times at age 89.

The importance of income deprivation is investigated further in the right-hand panel of Figure 1. In this figure we consider centiles rather than deciles, and use Age Standardised Mortality Rates (ASMRs) over ages 65 to 89 rather than crude age-specific death rates to dampen sampling variation. For an overview and the definition of ASMRs, see Supplementary Appendix B.

We see that income deprivation has a clear impact across all centiles, and, indeed, steepens towards both edges.

- At the right hand end (high deprivation and high mortality), the steeper curve might reflect the possibility that prolonged ill-health drives some people into more deprived areas.
- At the left hand end for the least deprived (which we interpret as most affluent), the rationale for a steepening curve is less clear, although (speculating) it might be that higher levels of wealth might facilitate better health care in old age. This pattern at the level of centiles can also be seen in US data (Chetty et al., 2016).

In Figure 2, we illustrate how the ASMRs for the deciles have changed over time for ages 65 to 89. This plot reveals the following features:

- We can see a widening gap between groups 1 and 10. This is more marked for females and for the older age group.
- Mortality improvements can be seen to have slowed down since 2010 or 2011. However, assessing the impact on deciles 1 to 10 needs care. The slowdown is more marked in this older age group than it is at younger ages. And it is also more marked in the most deprived groups, even after taking account of the fact that they had been experiencing slower improvements since 2001.

A clear takeaway from Figure 2 concerns the setting of future mortality improvement rates in applications such as population projections and actuarial valuations. Specifically, short-term improvement rates should be different for different socio-economic groups, with higher short-term improvement rates for the least deprived groups.

Lastly, we consider how mortality rates vary from region to region across England. In Table 1 (bottom row) we give the ratio of actual versus expected deaths by region using English national mortality for expected deaths. Corresponding values by income-deprivation decile are given in Table 1 (right-hand column) for comparison. We can see significant differences between regions leading to the well-publicised North/South divide. But we can also see that differences by region are dwarfed by differences between income-deprivation deciles.

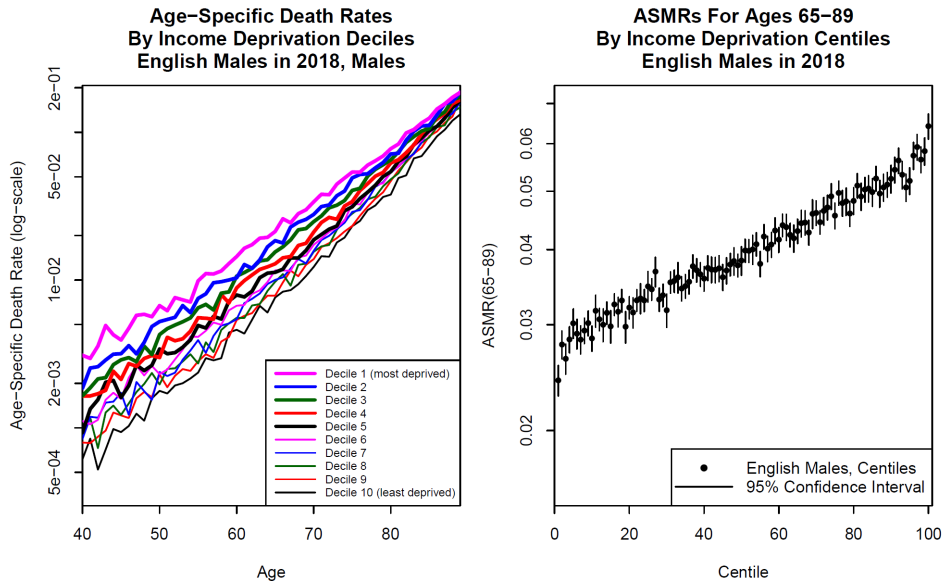

Figure 1: Left: Age-specific death rates for English males in 2018, by income-deprivation deciles. Right: Estimated ASMRs in 2018 by income-deprivation centile for ages 65 to 89. Centile 1: least deprived. Centile 100: most deprived. Bars show approximate 95% confidence intervals for ASMR estimates.

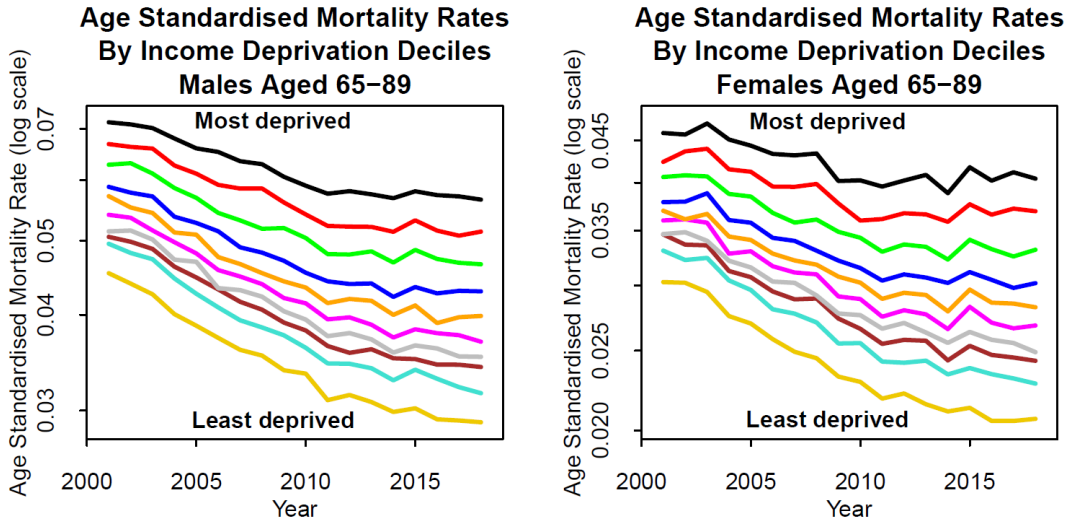

Figure 2: Estimated ASMRs by income-deprivation deciles over the period 2001 to 2018 for males and females ages 65 to 89.

	Region									
Decile	NE	NW	YH	EM	WM	E	L	SE	SW	All
1	157.3	159.6	151.6	147	143.3	138.3	133.4	139.4	134.9	**149.5**
2	139.5	144.5	134.9	134.6	133.9	123.8	127.7	129.3	125.9	**133.6**
3	132.2	128	122.8	121.7	126.5	112.8	114.9	119.8	108.6	**120.4**
4	119.5	120.6	111.7	112.2	108	102.9	105.2	110.7	98.4	**108.8**
5	105.7	107.4	105.4	104	98.1	97	96.3	102.5	89.9	**99.8**
6	104.4	99.1	97.7	98.3	94.9	88.6	91.2	93.8	88.1	**93.7**
7	95.3	92.1	92.6	91.3	90.9	86.5	82.1	87.7	82.7	**88.3**
8	89.4	90	90.6	87.9	86.3	81.8	78.5	85	80.8	**85.3**
9	83	82.9	82.9	82.2	80.6	77.5	77.4	78.7	76.1	**79.7**
10	72.9	74.3	71.9	74.9	72.3	71.5	68.6	71.6	67.5	**71.6**
All	*115.5*	*112.9*	*107.6*	*101.8*	*104.2*	*91.5*	*99.5*	*90.4*	*89.2*	**100**

Table 1: Ratios (%) of actual deaths to expected deaths by region and IMD decile for males aged 70 to 79 and 2001 to 2018. Expected deaths are based on national mortality and region/decile exposures by single year and single age. Final column (bold): actual to expected by IMD decile (no subdivision by region). Bottom row (italics): actual to expected by region (no subdivision by IMD decile). Regions: NE North East; NW North West; YH Yorkshore and the Humber; EM East Midlands; WM West Midlands; E East; L London; SE South East; SW South West.

4 A two-factor model with multivariate local linear regression

In this section we will outline the non-parametric method of *local linear regression* for assessing the relative mortality risk for a specific LSOA.

4.1 The model for LSOA-specific relative mortality risk

As defined in Section 2.2, let $m(t, a)$ to be the *crude* age-specific national death rate in year t, age a for one of the sexes. We now seek to model the *underlying* LSOA-specific death rates, $m(i, t, a)$, in LSOA i linking to a vector, $X(i)$, of predictive variables. The death rate satisfies the usual assumption that

$$E[D(i, t, a)] = m(i, t, a)E(i, t, a).$$

Our general model is then that, *over the limited age range* (a_0, a_1) *and range of years* (t_0, t_1),

$$m(i, t, a) = m(t, a)F_1(X(i))F_2(i)$$

where

- $F_1(x)$ is the relative risk due to the socio-economic (and other) characteristics of an LSOA with socio-economic characteristics x;
- $F_2(i)$ is a relative risk that captures remaining spatial effects once the $F_1(X(i))$ have been fitted.

Note that F_1 is a function of the vector of socio-economic characteristics x while F_2 is a function of an LSOA index i. That means the factor F_1 is defined for any characteristics x regardless of whether or not there is an LSOA with $X(i) = x$. The spatial factor F_2 could, of course, be defined for any geographical location but we think this is less important here since LSOAs divide England into rather small geographical units and can therefore themself be used to determine location.

We define the total number of deaths in LSOA i for a given age range, a_0 to a_1, and an observation period t_0 to t_1 as

$$D(i) = \sum_{t=t_0}^{t_1} \sum_{a=a_0}^{a_1} D(i,t,a).$$

Since the socio-economic and geographical factors, $F_1(x)$ and $F_2(i)$, are assumed to be constant over the given range of ages and years, we obtain for the expected number of deaths $E[D(i)]$ in LSOA i:

$$E[D(i)] = \sum_{t=t_0}^{t_1} \sum_{a=a_0}^{a_1} E[D(i,t,a)] = F_1(X(i))F_2(i)\hat{D}_0(i)$$

where

$$\hat{D}_0(i) = \sum_{t=t_0}^{t_1} \sum_{a=a_0}^{a_1} m(t,a)E(i,t,a).$$

In taking this approach,

- $\hat{D}_0(i)$ represents the baseline expected deaths with no allowance for socio-economic or other effects.
- $\hat{D}_0(i)F_1(X(i))$ represents our best estimate of the expected deaths based on socio-economic drivers alone.
- $\hat{D}_0(i)F_1(X(i))F_2(i)$ represents our best estimate of the expected deaths based on soecio-economic and spatial drivers.

4.2 Stage 1: estimate the socio-economic relative risk, F_1

To fit this model we, first, calculate the actual-over-expected death ratio based on the baseline model that there are no socio-economic effects:

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

We then use the $R_0(i)$ to derive estimates of $F_1(X(i))$ making use of the socio-economic predictive variable $X(i)$.

Unadjusted predictive variables take different ranges: some are on the scale $(0, 1)$, some $(0, 100)$, some $(-\infty, \infty)$. To aid comparison and computation of the local-linear-regression weights, the predictive variables are (unless stated otherwise) standardised as follows:

- Suppose that the LSOAs are indexed by $i = 1, \ldots, L$, and that the predictive variables are indexed by $j = 1, \ldots, n_P$.
- Let $P(i, j)$ be the unadjusted predictive variable.
- Define $X(i, j) = (P(i, j) - \mu_j)/\sigma_j$ where μ_j and σ_j are the empirical mean and standard deviation of $P(1, j), \ldots, P(L, j)$.

Hence, the $X(i, j)$ all have mean 0 and variance 1.

Variables that we chose not to standardise include

- Urban-Rural classification and region (which are categorical rather than ordinal variables);
- The proportions in communal establishments (including care homes with and without nursing) in a particular LSOA (which remain as proportions on a $(0, 1)$ scale).[4]

For any given vector x of socio-economic characteristics we now fit a linear regression function to all LSOAs giving a large weight to LSOAs j with $X(j)$ "close" to x. Thus, for a given x, we choose $\hat{a}(x)$ and $\hat{b}(x)$ such that they minimise

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

where the $w_1(x, j)$ are LSOA-specific weights related to the distance between $X(j)$ and x that are discussed below. We then set

$$\hat{F}_1(x) = \hat{a}(x) + \hat{b}(x)^T x, \tag{1}$$

and update the estimated deaths in any LSOA i as

$$\hat{D}_1(i) = \hat{D}_0(i)\hat{F}_1(X(i)) \tag{2}$$

and the actual-over-expected ratio

$$R_1(i) = D(i)/\hat{D}_1(i) = R_0(i)/\hat{F}_1(X(i)).$$

4.2.1 Stage 1 weights

The calculations above need the weights to be specified. Broadly speaking, the weight $w(x, j)$ is large if LSOA j has socio-economic characteristics, $X(j)$, that are similar to x. The weight tends to zero as $X(j)$ gets further away from x.

A number of different ways could be used to generate the weights. Here we start with the socio-economic distance between two vectors of characteristics x and y

$$d_1(x, y) = ||x - y||_S = \left(\sum_{k \in S} (x(k) - y(k))^2\right)^{0.5} \tag{3}$$

[4]From a statistical perspective, and in this analysis, care homes can be considered to be nuisance parameters. Specifically, the presence of a care home in an LSOA should not impact on the socio-economic characteristics and mortality experience of the non-care-home population in the same LSOA.

where S is a subset of $(1, \ldots, n_P)$. The reason why S might not be the full set of predictive variables $1, \ldots, n_P$, is to allow for the exclusion of "nuisance" variables such as care-home proportions that do not contribute to the underlying socio-economic characteristics of an LSOA but do influence mortality rates. The urban-rural class is also not included in S but will play a role in the final definition of the weights. We next rescale the distances

$$v_1(x, y) = d_1(x, y)/d_{1\max}(x)$$

where we choose to use an adaptive bandwidth, $d_{1\max}(x)$. Lastly, we define the weights as

$$w_1(x, j) = \begin{cases} \hat{D}_0(j)\left(1 - v_1(x, X(j))^3\right)^3 & \text{if } v_1(x, X(j)) \leq 1 \text{ and } u(X(j)) = u(x) \\ 0 & \text{if } v_1(x, X(j)) > 1 \text{ or } u(X(j)) \neq u(x) \end{cases}$$

where $u(x)$ is the urban-rural class decoded in x: that is, we give zero weight to LSOAs that are not in the same urban-rural class.

When we are estimating the parameters, we need to be mindful of the fact that for some LSOAs, i, the position of $X(i)$ lies in a relatively dense region while for other LSOAs $X(i)$ lies in a relatively sparse region. If we use a constant bandwidth, then, in the latter case, relatively few neighbours of $X(i)$ will contribute to the estimation of the parameters leading to a high standard error. To avoid this we use an adaptive bandwidth, $d_{1\max}(i)$, so that the weight function for LSOAs in sparse regions includes data from a sufficiently large number of neighbours. In our case, we have chosen $d_{1\max}(i)$ to equal the distance to the K_1th nearest neighbour in the same urban-rural class: that is, only the $K_1 - 1$ nearest neighbours in the same urban-rural class will have a non-zero weight, with the weights diminishing to zero the further they are away from $X(i)$. More specifically, we have chosen $K_1 = 2500$.

There are two further elements in the weight function: the commonly used tricube kernel function; and the expected deaths, $\hat{D}_0(j)$. The purpose of scaling the weights by $\hat{D}_0(j)$ is to give greater weight to LSOAs that have larger numbers of people (and, hence, larger numbers of deaths) in the age range of interest: larger numbers means that sampling variation in the actual-over-expected variable $R_0(j)$ for a particular LSOA is lower. This contrasts with, and generalises, the standard assumption in non-linear regression that all individual observations have the same variance.[5]

4.3 Stage 2: estimate the remaining spatial relative risk

In this second stage, we assess the remaining spatial relative risk based on the updated actual-over-expected, $R_1(i)$, using only the longitude and latitude, $Y(i) = (Y_1(i), Y_2(i))$, of each LSOA, i, and the urban-rural class.

Unlike the socio-economic predictive variables, there is no *ex-ante* reason why we might expect any systematic trends in relative risk as we move from west to east or north to

[5]To justify this particular scaling, suppose we have a sequence of random variables $Y_j \sim N(\mu, \sigma_j^2)$ where the σ_j are known and μ has to be estimated. An estimator for μ is $\sum_j w_j Y_j$ with $\sum_j w_j = 1$. The estimator with the lowest variance has $w_j \propto 1/\sigma_j^2$. In our case the observations $R_0(j)$ have a variance that is proportional to $1/\hat{D}_0(j)$ (in the absence of socio-economic effects).

south. We will, therefore use the simpler method of kernel smoothing to generate estimates of the relative risk:

$$\hat{F}_2(i) = \frac{\sum_j w_2(i,j) R_1(j)}{\sum_j w_2(i,j)}$$

where the weights depend on the physical distance between LSOAs i and j, with adjustment for the urban-rural class of each.

4.3.1 Stage 2 weights

These are based on physical distance between LSOAs adjusted for urban-rural class.

$$\begin{aligned} d_2(i,j) &= \left[\frac{(Y_1(i) - Y_1(j))^2 + (Y_2(i) - Y_2(j))^2)}{\phi(u(i))\phi(u(j))} \right]^{0.5} \\ v_2(i,j) &= d_2(i,j)/d_{2\max}(i) \\ w_2(i,j) &= \begin{cases} \hat{D}_1(i)\left(1 - v_2(i,j)^3\right)^3 & \text{for } v_2(i,j) \le 1 \\ 0 & \text{for } v_2(i,j) > 1 \end{cases} \end{aligned}$$

As with the $w_1(i,j)$ we allocate greater weight to LSOAs with higher expected deaths. The $\phi(u)$ scaling parameters are dependent on the urban-rural class $u = 1, \ldots, 5$. The rationale for this feature is that, e.g., rural LSOAs tend to be much further apart than urban LSOAs. Thus a rural neighbouring LSOA 10km away might carry more weight than a city-based LSOA 1km away. Here, we have used $\phi = (1.5, 1.5, 5, 25, 1)$ and the adaptive bandwidth, $d_{2\max}(i)$, is set equal to the distance to the K_2th nearest neighbour (based on the $d_2(i,j)$ distances), with $K_2 = 250$ giving satisfactory results.

4.4 Advantages and disadvantages of local linear regression

Amongst non-parametric methods, a key reason for using local linear regression (LLR) for estimating the socio-economic factor F_1 rather than, for example, kernel smoothing is that it captures explicitly the local slope in the data. This is important where the response variable (here, the relative risk) is believed to be increasing or decreasing as the underlying predictive variables change. This is specifically useful when we estimate $F_1(x)$ for any x whose (socio-economic) neighbours $X(j)$ are mostly to one side rather than evenly distributed around x: for example, if x is near the edge of the data.

Second, LLR offers an approach that can be easily implemented in multiple dimensions.

A third advantage is that estimates of relative risk are not especially sensitive to non-linear transformations of the predictive variable, x (e.g. a log-transform). The exception to this might be at the edges of the data when those points with an appreciable weight, $w(x,j)$, are spread out over a wider range of values of x. This contrasts, for example, with Generalised Linear Models (GLMs), where, by design, the (transformed) expected relative risk is linear in the predictor x: so x needs to be scaled and transformed accordingly before model fitting to get a good fit.

Fourth, LLR automatically captures interactions between different predictive variables in multiple dimensions. With GLMs, interactions must be investigated in a systematic way, and this could be very challenging computationally in multiple dimensions.

	Predictive Variable	Scaling
1	Income deprivation (older people)	standardised
2	Employment deprivation	standardised
3	Average number of bedrooms	standardised
4	Age 65+ with higher education	standardised
5	Proportion born in the UK	standardised
6	Overcrowding in housing	standardised
7	Affordability problems	standardised
8	Poor air quality	standardised
9	No central heating	standardised
10	care home *with* nursing, 60+	$[0, 1]$ proportion
11	care home *without* nursing, 60+	$[0, 1]$ proportion
12	urban-rural class	categories 1 to 5

Table 2: Predictive variables and scalings used in the core analysis. (See Supplementary Appendix A for a more detailed description of the predictive variables.)

The main disadvantage of LLR and other non-parametric methods is the lack of easily interpretable parameters: results cannot be communicated in a simple way through, for example, the specification of a simple function. This is in contrast to parametric models like GLM for which each estimated parameter can usually be given an interpretation. Nevertheless, LLR does provide detailed outputs such as the local slope, $b(x)$, which allows users to explore the influence of individual parameters across different sections of the data.

Many of these advantages and disadvantages are shared with the random forest (RF) algorithm (Wen et al., 2023) and the two methods produce similar results. RF is computationally less intensive than LLR. On the other hand, LLR should, in general, perform better than RF near the edges of the data, and local estimates of the slope, $b(x)$, in LLR allow easier exploration of the results.

5 Results

5.1 The core model

We now present results for our core model, using the data from 2001 to 2018. We experimented with different groups of predictive variables and settled on those listed in Table 2 on the basis of (a) lower variance of randomised probability-transformed residuals (see Supplementary Appendix D) and (b) lower variance of the remaining spatial relative risk. In aiming to minimise the remaining spatial relative risk *we are seeking to explain as much as possible of the observed variation in mortality using socio-economic variables.*

The zeros in the weight function, $w_1(i, j)$, mean that the model for each urban-rural (UR) class is fitted independently of other urban-rural classes. Within each UR class we have 11 predictive variables, out of which care home proportions are to be treated as nuisance parameters rather than socio-economic parameters. Hence, in equation (3),

$S = \{1, 2, \ldots, 9\}$.

The analysis produces 32,844 estimates each for the $F_1(X(i))$ and $F_2(i)$. Since we only consider the socio-economic factor F_1 for characteristics observed in an LSOA, we will use the short-hand notation $F_1(i)$ for $F_1(X(i))$ in what follows. Empirical distributions (CDFs) for these are plotted in Figure 3 for males. Results for females are very similar. The two plots show CDFs for each of the five age groups. First, consider the left hand plot: socio-economic relative risk, $F_1(i)$. We can see that the wide spread for the age 40-49 age group gradually narrows with age (consistent with Figure 2). For ages 40-49, the plot reveals how large the mortality inequality gap is between the top and bottom groups: more than 7% of LSOAs have a death rate that is less than half of the national rate, while 6% have mortality that is more than double the national rate. The factor of $4\times$ is approximately equivalent to an effective age difference of 14 years.

Second, consider the right-hand plot: the remaining spatial relative risk, $F_2(i)$. The most striking feature of these CDFs is that they are much narrower than the CDFs for socio-economic risk. Indeed, the empirical variance of $F_2(i)$ is in the range 1.4% to 3.0% of the variance of $F_1(i)$ for the same age group. In part, this reflects our desire to explain as much of the variation in mortality using socio-economic predictive variables first, and spatial variation second. However, it indicates that, with the right choice of socio-economic variables, socio-economic information by far outweighs geographical location as a predictor of mortality.

Figure 4 (males, ages 40-49) maps the estimated values of the socio-economic relative risk, $\hat{F}_1(i)$, (left), the spatial relative risk, $\hat{F}_2(i)$ (middle), and the combined relative risk, $\hat{F}_1(i)\hat{F}_2(i))$ (right), plotting values using a limited colour range.

The middle plot shows how $\hat{F}_2(i)$ varies across England. We can see that there are patches of oranges/reds and greens/blues (higher and lower mortality than predicted by the socio-economic model only), but there is no systematic north/south or other divide. The reasons for these patterns is currently not clear but they will, amongst other things, reflect additional behavioural variations (e.g. spatial variation in smoking prevalence) that are not well represented in our existing set of predictive variables, including those in our wider dataset.

The left and right-hand plots in Figure 4 show the socio-economic-only (left) and combined (right) relative risk, $\hat{F}_1(i)$ and $\hat{F}_1(i)\hat{F}_2(i)$. We can see small differences between the two maps, but essentially they are very similar, reflecting our earlier observation that the residual spatial risk is very much less variable than the socio-economic risk. Across the map, there is apparently much more blue, but blues tend to be in less densely populated LSOAs, sometimes rural, and therefore much larger in area. The oranges and reds, representing high relative risk, tend to be in high-density inner-city areas in London, and northern cities.

Equivalent plots for the higher age groups and for females can be seen in the supplementary Supplementary Appendix E. In terms of remaining spatial relative risk, there is is a gradual shift in the geographical distribution as well as a narrowing of the range of values and the emergence of a split in the form of south coast and east (low) versus the rest of the country, and persistent higher mortality in west London. For the combined relative risk the pictures are more consistent with Figure 4 but with a gradually diminishing range, consistent with Figure 3.

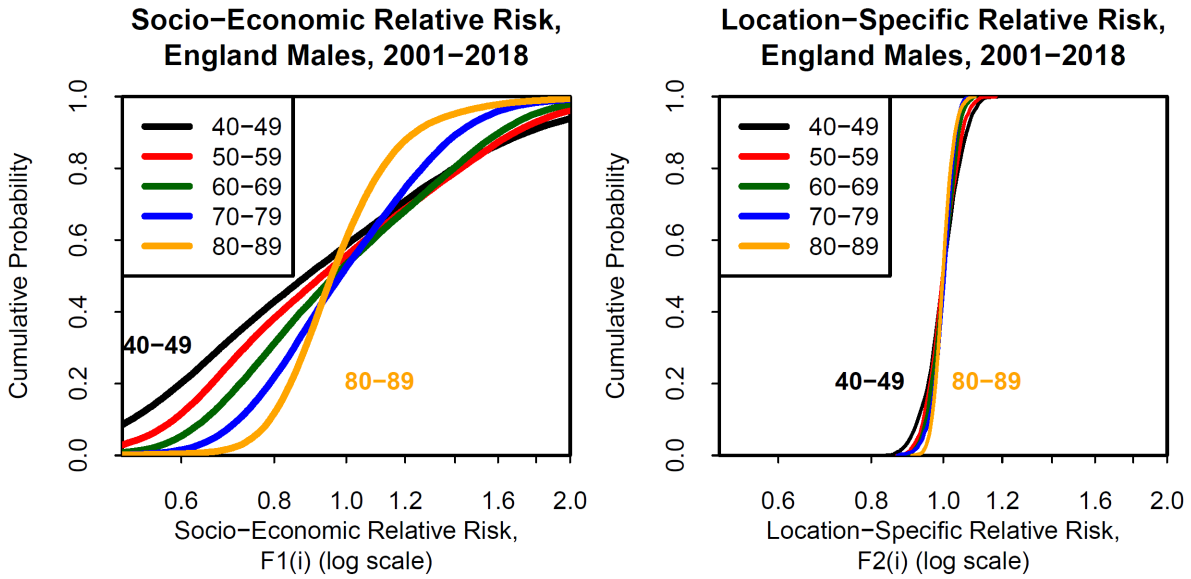

Figure 3: Left: Empirical cumulative distribution functions for the estimated socio-economic relative risk $F_1(X(i))$ for English males, age groups 40-49, ..., 80-89, each covering years 2001-2018. Right: Empirical cumulative distribution functions for the estimated remaining spatial relative risk $F_2(i)$ for age groups 40-49, ..., 80-89.

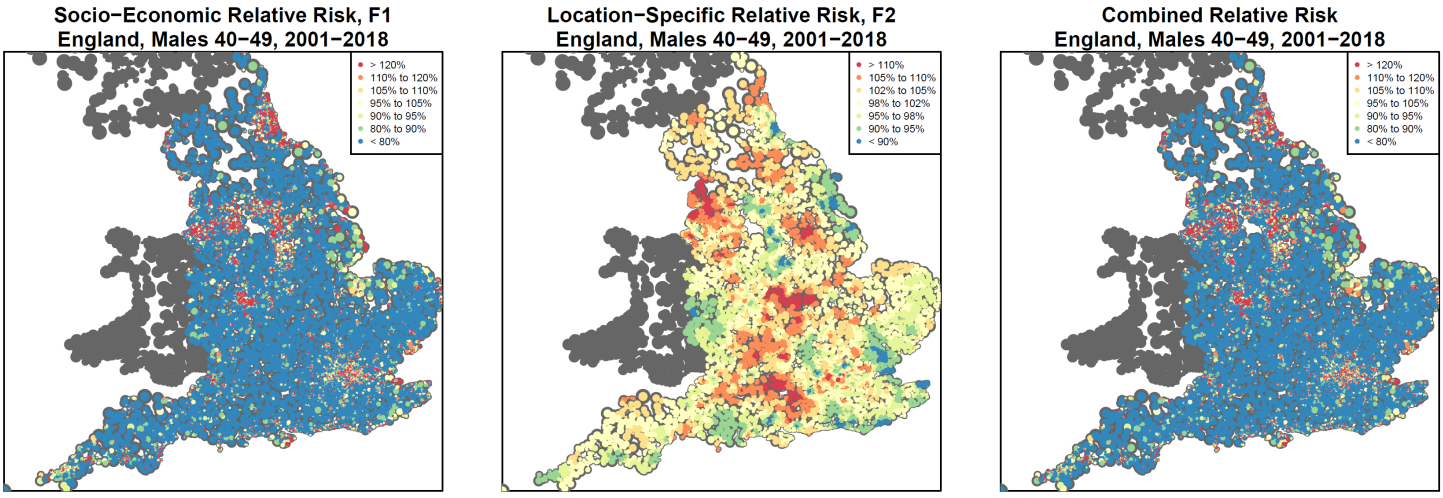

Figure 4: Estimated relative risks by LSOA for English males aged 40-49. Left: estimated socio-economic relative risk (coloured dots), $F_1(i)$. Middle: estimated spatial relative risk, $F_2(i)$. Right: combined relative risk, $F_1(X(i))F_2(i)$. Dot sizes reflect the physical size of each LSOA.

5.2 Care homes

The presence of a care home in an LSOA can clearly have an impact on observed mortality. For example, an affluent LSOA that we would expect to have low mortality might have higher than expected deaths if there is a large care home in the neighbourhood. As remarked above, therefore, care homes with and without nursing have been included as predictive variables, but, as they are regarded as nuisance variables, they do not influence the weight function, $w_1(i, j)$.

In the data, there are 23,464 LSOAs with no care home, 6,023 have care homes without nursing only, 2,058 have care homes with nursing only, and 1,299 have care homes of both types. In 2011, there were 177,530 persons above age 60 in a care home without nursing[6] and 131,158 persons in a care home with nursing.[7]

Care home data are available for each LSOA (proportions in the 60+ age group in care homes with and without nursing). It is therefore of interest to consider the following questions: *are care homes concentrated more in either less or more deprived areas; can we estimate mortality rates for the healthier non-care-home population; what do mortality rates look like if each LSOA has an average sized care home population?*

To look at these questions we consider three variants of the socio-economic relative risk, $F_1(i)$.

- Case A: this is our base case using Equation (1). We label this as $\hat{F}_1^A(i)$ and it provides an estimate of the socio-economic relative risk using the actual care home proportions.
- Case B: we assume that none of the population is resident in a care home with or without nursing, that is, we estimate the socio-economic factor for LSOA i using the estimated coefficients $\hat{a}(i)$ and $\hat{b}(i)$, see Equation (1), but we modify the vector of predictive variables, $X(i)$, to set the care home populations to zero:

$$\hat{F}_1^B(i) \;=\; \hat{a}(i) + \sum_{j=1}^{9} \hat{b}(i,j)X(i,j) + \hat{b}(i,10)\cdot 0 + \hat{b}(i,11)\cdot 0.$$

- Case C: we assume that the average proportions $\bar{X}_{10}$ and $\bar{X}_{11}$ of the 80-89 population reside in a care home with and without nursing,

$$\hat{F}_1^C(i) = \hat{a}(i) + \sum_{j=1}^{9} \hat{b}(i,j)X(i,j) + \hat{b}(i,10)\bar{X}_{10} + \hat{b}(i,11)\bar{X}_{11}.$$

To address the first question above, we investigated the proportion in a care home versus what we estimate to be the relative risk for the non-care-home population, $\hat{F}_1^B(i)\hat{F}_2(i)$. (See Figure 23, right-hand panel, in Supplementary Appendix E.) We found that care homes seem to be evenly distributed across all LSOAs with no bias towards areas of high or low mortality.

[6] Average 18 persons in those LSOAs with a care home without nursing.

[7] Average 53 persons in those LSOAs with a care home with nursing.

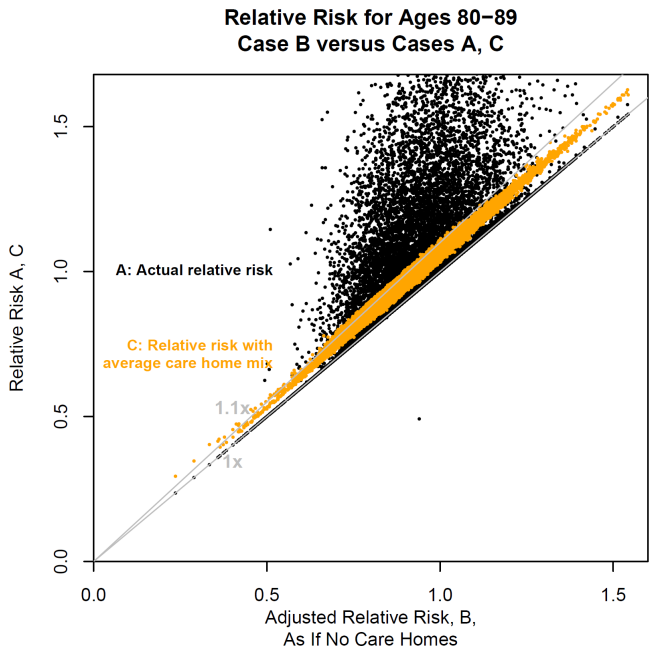

Figure 5: Scatterplot of the adjusted relative risk $\hat{F}_1^B(i)\hat{F}_2(i)$ (case B) versus cases A and C. A (black dots): actual relative risk $\hat{F}_1^A(i)\hat{F}_2(i)$. C (orange dots): relative risk assuming average proportion in care homes with and without nursing $\hat{F}_1^C(i)\hat{F}_2(i)$. Diagonal lines (gray lines) with gradients 1 and 1.1 are included to aid interpretation.

The second and third questions use the adjusted relative risks, $\hat{F}_1^B(i)$ and $\hat{F}_1^C(i)$. This is illustrated in Figure 5 where, for ages 80-89, we compare relative risks with and without inclusion of the care home effect (Case A versus Case B in the figure, black dots). All LSOAs with no care homes lie on the 1x diagonal. Those with a care home lie above the gray 1x diagonal. The ratio with to without is affected by two factors: the number of people in a care home with or without nursing; and the estimated "impact" on individuals of being in a care home (the magnitude of the $\hat{b}(i, 10)$ and $\hat{b}(i, 11)$ parameters). As we can see, in some LSOAs the presence of care homes more than doubles the estimated mortality rate within an LSOA. Cases B and C are also compared in the Figure 5 (orange dots). In case C, the impact of having the average care home population is somewhere between a parallel shift and a proportional adjustment to the adjusted case with no care homes. In aggregate, case C (as with case A) should correspond approximately to national mortality, whereas case B will have lower aggregate mortality than the national population as it excludes the effect of care homes.

5.3 Sensitivity of estimated relative risk to predictive variables

By way of example, Figure 6 illustrates how the relative risk with care homes removed, $F_1^B(i)$, depends on given pairs of predictive variables. Individual points (LSOAs) are coloured according to which decile $F_1^B(i)$ falls in.[8] The left-hand plot focuses on males

[8]The extent to which the coloured bands are fuzzy or lack crisp divisions is an indication that the two predictive variables do not, on their own, provide a complete picture: there is additional information in the other variables.

aged 70-79 and looks at the joint impact of income deprivation amongst the old and educational attainment (the proportion with a BSc or higher) amongst the elderly. For low and high levels of income deprivation, the level of educational attainment does not seem to have much impact. But where there are moderate levels of income deprivation (e.g. 30%) the level of educational attainment can be seen to be much more significant with, everything else being equal, higher levels of educational attainment being associated with lower mortality. This, therefore, illustrates an interaction between the predictive variables that the LLR method picked up without the need to be pre-specified. In the right-hand plot we see the impact on males aged 50-59 of employment deprivation and overcrowding. Here we can see that both predictive variables have a consistent effect with higher values predicting higher mortality and employment deprivation having a rather stronger impact.

The left hand plot in Figure 7 shows how the relative risk for males aged 70-79 depends on imcome deprivation amongst the elderly and the proportion born in the UK. 85% of LSOAs have a proportion of UK born above 70%, and, above this threshold of 70%, relative risks seem to have only a weak dependence on the proportion of UK born. However, below this threshold the decile bands become more diagonal indicating that there is a clear dependency between relative risk and the proportion of UK born. Most LSOAs below this threshold belong to urban conurbations, especially London, so the dependency is only evident in urban-rural classes 1 and 5. This dependency is confirmed when we look at the results of the fitted local regressions in the right-hand plot. Here we plot $\hat{b}(i,5)$ against $X_5(i)$, the proportion of UK born. Although there is a fair degree of noise in the parameter estimates, the $\hat{b}(i,5)$ are clearly positive for urban-rural classes 1 and 5 (black and orange dots) respectively below about 70%. A formal analysis of this effect is beyond the scope of this paper, but three possible explanations come to mind. First, if the ethnic mix of an LSOA with a high proportion of first-generation immigrants is quite homogeneous then there the elderly might benefit from a stronger community-wide support network or family (for example, in multi-generation households) than is present in neighbourhoods where the proportion of UK born is higher. Second, even at these advanced ages, there might be a significant healthy-immigrant effect (see, for example, Vang et al., 2017, and Wen et al., 2020). Third, further analysis indicated that the LSOA's with a low proportion of UK born also have relatively low employment deprivation compared to levels of old-age income deprivation. This suggests that the elderly population of first-generation immigrants might have been active but low-paid workers who need income-support in retirement.

At lower ages, the proportion of UK born has a more-complex but still-significant pattern of impact. For example, in London, in the 40-49 age group, the pattern discussed above reverses: everything else being equal, the lowest proportions of UK born are associated with higher mortality. Again this needs further investigation and the underlying reasons might be more complex still.

5.4 Regional and decile mortality

In Table 1 earlier, we noted how much variation there was in regional mortality versus national mortality. We now compare expected deaths at the regional level before and after adjusting for the socio-economic relative risk $\hat{F}_1(i)$. These are reported in Table

Figure 6: Scatterplots showing the relationship between pairs of predictive variables and socio-economic relative risk, $F_1^B(i)$, with the care home effect removed. Left: males aged 70-79, income-old deprivation versus proportion amongst the old with a high education. Right: males aged 50-59, employment deprivation versus overcrowding. Dots: individual LSOAs. Colours: dark brown, decile 1, 10% highest relative risk, $F_1^B(i)$; dark blue/green, decile 10, 10% lowest relative risk.

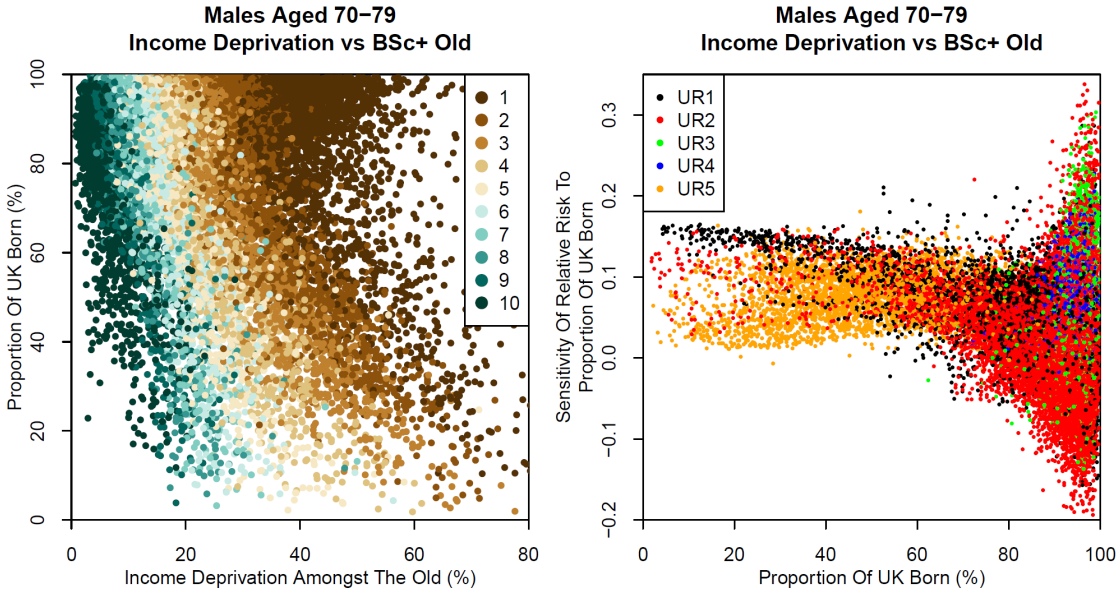

Figure 7: Left: As Figure 6 but showing the dependence of the relative risk for males aged 70-79 on income deprivation amongst the old and the proportion born in the UK. Right: Scatterplot showing the sensitivity of the relative risk to changes in the proportion of UK born coloured by urban-rural class: that is, $\hat{b}(i,5)$ against $X_5(i)$ respectively.

	Regional Relative To National Mortality			
	Males		Females	
Region	Unadjusted	Adjusted	Unadjusted	Adjusted
North East	115.5	100.8	120.5	101.6
North West	112.9	102.5	116.1	103.6
Yorkshire and The Humber	107.6	100.4	108.3	100.2
East Midlands	101.8	100.4	102.1	99.4
West Midlands	104.2	98.9	102.6	97.3
East	91.5	97.2	92.5	97.7
London	99.5	101.1	95.2	101.0
South East	90.4	100.6	89.3	100.1
South West	89.2	98.1	87.3	97.2

Table 3: Comparison of regional death counts over the age range 70 to 79, and 2001-2018 relative to expected deaths using English national mortality, without (unadjusted) and with (adjusted) allowance for socio-economic relative risk, $F_1^A(i)$.

3. As can be seen, much of the regional variation that we observed previously (Table 3, males/females "unadjusted") can be explained by socio-economic effects (males/females "adjusted"). However, some differences still remain (but now much reduced) with the East, South West and North West standing out. Notably, the West Midlands shifts from having higher than expected mortality to slightly lower, while for London the reverse is true for females.

In Figure 8 we plot Age Standardised Mortality Rates (ASMRs) for males for the age range 65-89 based on two sets of deciles:

- deciles based on income deprivation (ID);
- deciles based on the estimated relative risk with the impact of care homes removed, $\hat{F}_1^B(i)$, using the 70-79 age group.[9]

We can see that the two plots have broadly similar patterns: improving mortality and a widening gap. However, the $\hat{F}_1^B(i)$-based deciles produce a slightly wider spread, consistent with greater predictive power.

In Figure 9 we plot Age and Deprivation Standardised Mortality Rates (ADSMRs) where we standardise using either income deprivation deciles or deciles based on the socio-economic mortality $\hat{F}_1^B(i)$.[10]

We can make the following observations:

- The unadjusted ASMRs by region (left-hand plot) highlight the significant differences in mortality between regions. This is partly mitigated by the deprivation

[9] $\hat{F}_1^B(i)$ is used, first, to map each LSOA to a decile. ASMRs can then be calculated for each decile over any age range (e.g. 65-89) using the raw deaths and exposures data.

[10] See Supplementary Appendix B.3 for definitions and a description of ADSMRs, and how ASDMRs mitigate the differences between ASMRs that are simply caused by the fact that some regions are more deprived than others.

adjustment based on income deprivation (middle plot).

- Use of the socio-economic-mortality deciles (right-hand plot) results in a much narrower range of ADSMRs than the income deprivation deciles: again an indication that the new approach in this paper explains much more of the differences between regions compared to income deprivation as a sole predictive variable. If London is excluded, then the gap between regions is almost halved.
- London is a clear outlier amongst the nine regions with much stronger improvements over the 18 year period than all other regions.
- All regions have seen a slowdown in mortality improvements since 2010/11. For London, this is a slowdown relative to its previously faster rate of improvement.
- For females (not plotted) the patterns are quite similar, but rates are lower, although regional differences are a bit bigger.

Explaining the London effect is beyond the scope of this paper, but it clearly needs to be better understood. Possible reasons are: changing demographics/gentrification[11]; a widening gap in spending on the National Health Service (NHS); more effective use of NHS and public-health funding in London than elsewhere. Note, also, that the predictive variables are single values covering the whole of the period 2001-2018 rather than time dependent, so it is possible that any drift in the predictive variables (e.g. income or employment deprivation) over time that is significantly different in London from other regions might explain the London effect.

5.5 Time dependency

It is of interest to investigate how much the relative risks vary through time. Here, we do this by fitting the model again to the non-overlapping time periods 2001-2009 and 2010-2018.

Figure 10 shows how the socio-economic relative risk changes between the two periods. As expected, there is a very high correlation between estimates for the two periods. The scatterplots also highlight differences between the different urban-rural classes.

- For the 40-49 age group, most points lie close to the main diagonal. This is an indication that levels of inequality have remained roughly similar. Much of the "fuzziness" around the main diagonal will be due to sampling variation in the death counts. However, for London (orange dots) the dots clearly lie on a flatter slope, and this is an indication that levels of inequality in this age group have *fallen* between the two periods in London.
- For the 80-89 age group, the scatterplot is clearly steeper than the main diagonal: a clear indication that levels of inequality have risen sharply in this age group in recent years.

[11] E.g. faster growth in London than elsewhere of GDP, or higher levels of education, or patterns of migration within England and from outside its borders might have benefitted London mortality (e.g. the *healthy immigrant effect*; see Vang et al., 2017, and Wen et al., 2020).

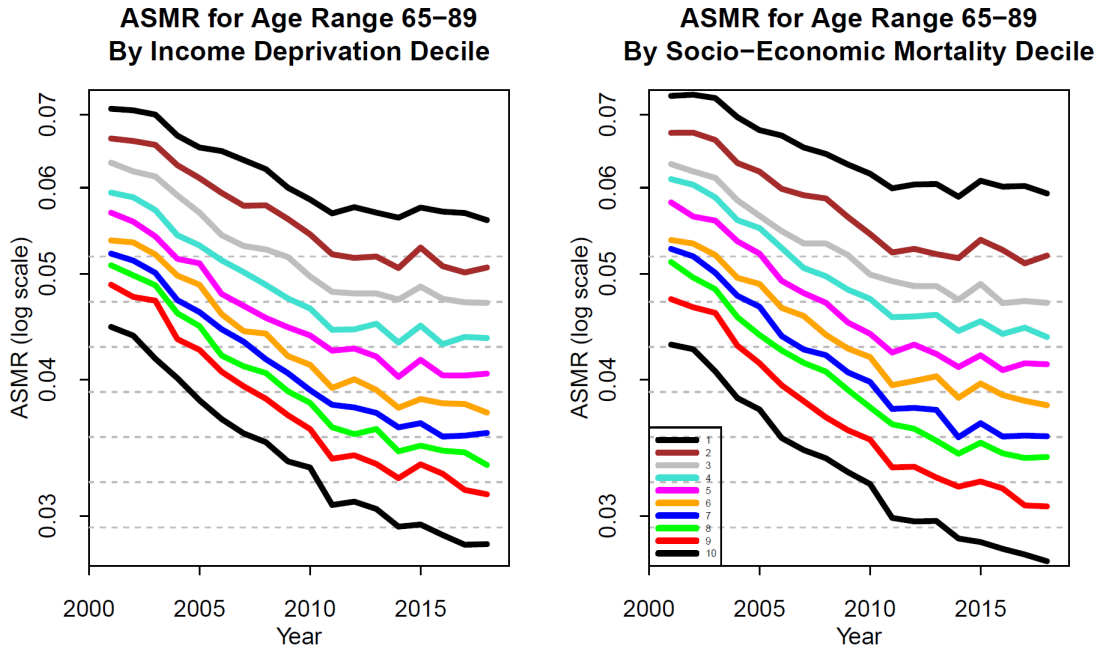

Figure 8: ASMRs for males for the age range 65-89 for each of 10 deciles. Left: deciles based on income deprivation deciles. Right: deciles based on the estimated socio-economic relative risk with the effect of care homes removed, $\hat{F}_1^B(i)$, for the 70-79 age group.

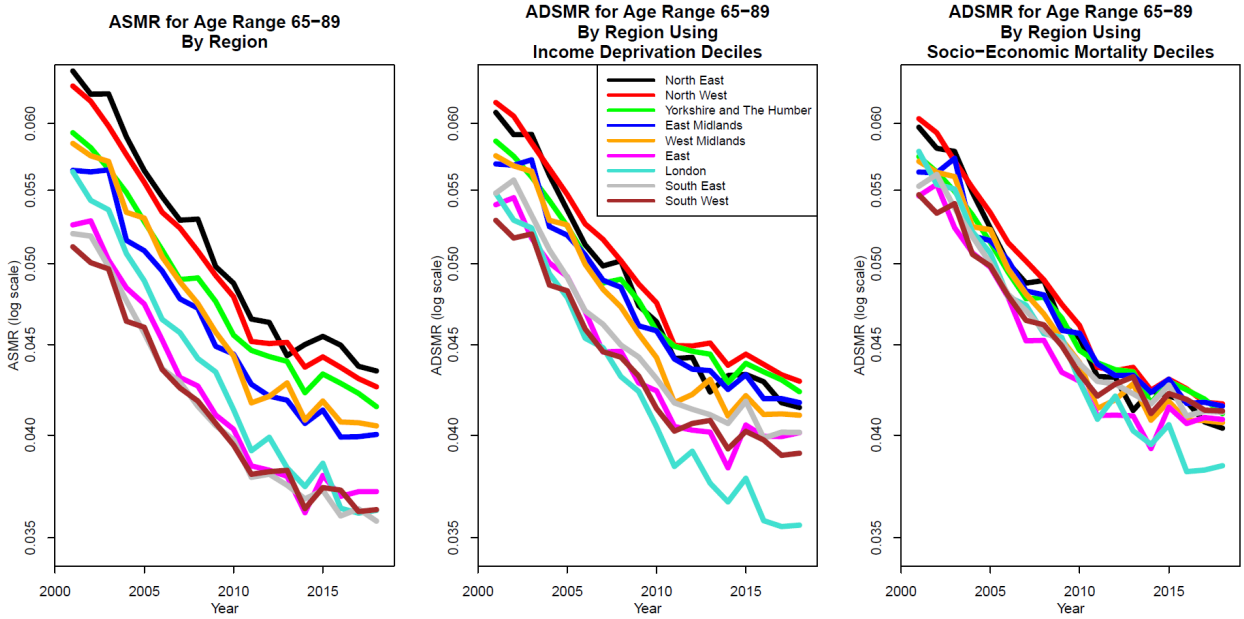

Figure 9: ASMRs and ADSMRs for males for the age range 65-89 for each region. Left: Unadjusted for variations in deprivation. Centre: deprivation standardisation using income deprivation. Right: deprivation standardisation using the estimated relative risk $\hat{F}_1^B(i)$ for the 70-79 age group.

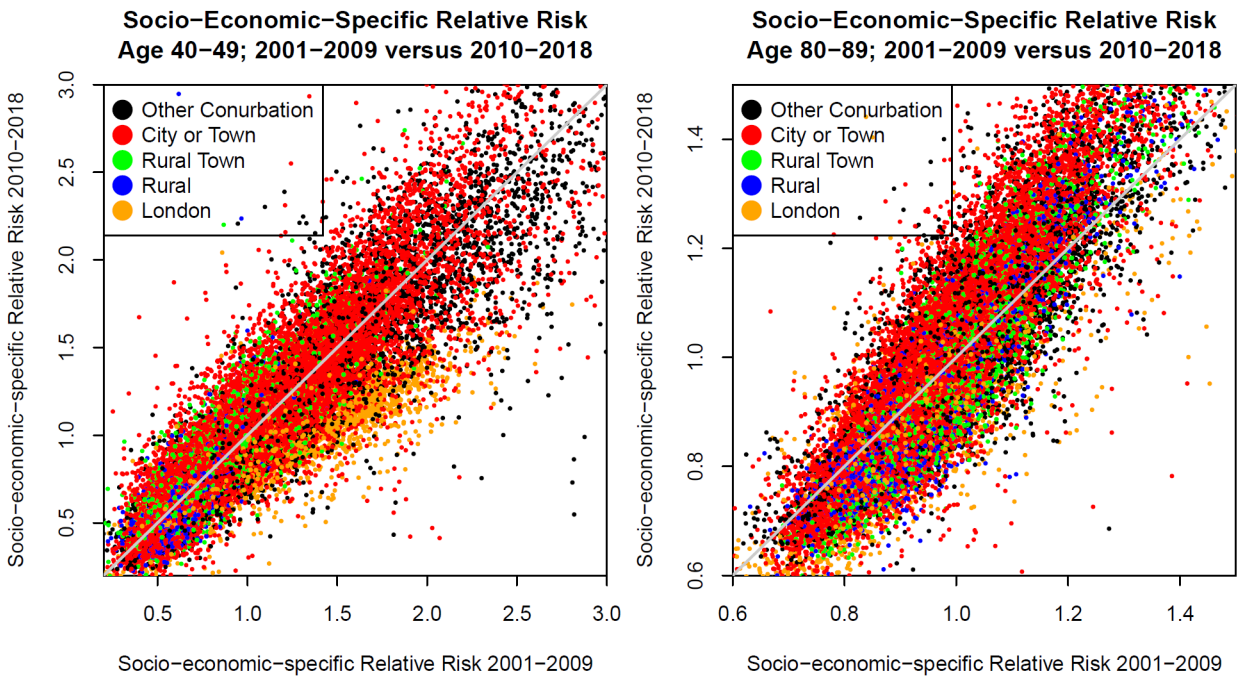

Figure 10: Socio-economic relative risk for males, ages 40-49 and 80-89. 2001-2009 estimates versus 2010-2018 estimates.

6 Conclusions

The increasing availability of large mortality-related datasets opens up the possibility for more detailed analyses of the key drivers of death rates.

In this paper, we have conducted a detailed analysis of mortality inequalities in England, using all-cause mortality data at the level of LSOAs. We have used the non-parametric method of local linear regression to quantify more accurately the very significant mortality inequalities that exist across England, particularly at younger ages. The method is very well suited to our large dataset and can handle, in a straightforward way, the inclusion of several predictive variables, and is not particularly sensitive to non-linear transformations of the predictive variables. In particular, the method automatically captures any potential interactions between predictive variables.

Amongst all of the available predictive variables, old-age income deprivation and employment deprivation were found to have the strongest predictive power. But we also found that urban-rural class and the presence of care homes within a neighbourhood were important predictors as well. Once socio-economic effects have been filtered out, the remaining spatial relative risk was found to be quite small in comparison, countering the headline differences between English regions. Perhaps this is not surprising, but it emphasizes that, on a like for like basis, there is no intrinsic disadvantage in terms of life expectancy to living in the north rather than the south.

The data also confirm that inequality between different socio-economic groups has been rising including at high ages where, in the past, there has been a generally narrower inequality gap compared to younger ages.

Our estimates of the relative risk (with and without care-home adjustments) can potentially be used in three ways. First, the estimated relative risk with the effects of care homes removed can be used to group LSOAs into deciles, with a clear improvement over the IMD or income deprivation as a predictor of high or low mortality. Second, the relative risk estimates can be used as predictive variables in their own right (on a continuous scale) in the assessment of the mortality of life insurance and pensions portfolios alongside other predictive variables such as pension amount and geodemographic grouping. Third, the socio-economic relative risk estimates might also be of use in developing strategies to tackle mortality inequality by identifying more accurately the worst affected groups. The companion spatial relative risks offer additional insights for policymakers into which areas have excess mortality even after taking account of socio-economic variation, potentially a result of poorer health behaviours in certain regions such as smoking.

Acknowledgements

The authors gratefully acknowledge funding from the Actuarial Research Centre of the Institute and Faculty of Actuaries, the Society of Actuaries and the Canadian Institute of Actuaries through the "Modelling Measurement and Management of Longevity and Morbidity Risk" research programme (see `www.actuaries.org.uk/arc`).

This study is also part of the research programme at the Research Centre for Longevity Risk - a joint initiative of NN Group and the University of Amsterdam, with additional funding from the Dutch government's Public Private Partnership programme.

References

Cairns, A.J.G., Kallestrup-Lamb, M., Rosenskjold, C.P.T., Blake, D., and Dowd, K., (2019) Modelling Socio-Economic Differences in the Mortality of Danish Males Using a New Affluence Index. *ASTIN Bulletin*, 49: 555-590.

Cairns, A.J.G., Blake, D., Dowd, K., and Kessler, A.R., (2016) Phantoms Never Die: Living with Unreliable Population Data. *Journal of the Royal Statistical Society, Series A*, 179: 975-1005.

Case, A., and Deaton, A. (2015) Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century. *Proceedings of the National Academy of Sciences*, 112: 15078-15083.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A, and Cutler, D. (2016) The association between income and life expectancy in the United States, 2001-2014. *Journal of the American Medical Association,* 315: 1750-1766.

Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74: 829–836.

Eurostat (2013) *Revision of the European standard population. Report of Eurostat's task force. 2013 edition.* Luxembourg: Publications Office of the European Union.

Longevity Science Panel (2018) Life expectancy: Is the socio-economic gap narrowing?

www.longevitypanel.co.uk/viewpoint/life-expectancy-is-the-socio-economic-gap-narrowing/ (Accessed 10/12/2019)

Macdonald, A.S., Richards, S.J., and Currie, I.D. (2018) *Modelling mortality with actuarial applications*. Cambridge University Press, Cambridge.

Mackenbach, J.P., Bos, V., Andersen, O., Cardano, M., Costa, G., Harding, S., Reid, A., Hemström, Ö., Valkonen, T., and Kunst, A.E. (2003). Widening socio-economic inequalities in mortality in six Western European countries. *International Journal of Epidemiology,* 32: 830-837.

Mackenbach, J.P., et al. (2016) Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 European countries. *Journal of Epidemiology and Community Health,* 69: 207–217.

Office for National Statistics (2015) English indices of deprivation 2015. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 (Accessed 3/11/2019)

Office for National Statistics (2018) Population estimates by output areas, electoral, health and other geographies, England and Wales: mid 2017. *Statistical Bulletin*, 25 October 2018.

Redondo Lourés, C., and Cairns, A.J.G. (2020) Mortality In The US By Education Level. *Annals of Actuarial Science*, 14: 384-419

Richards, S. J. (2008) Applying Survival Models to Pensioner Mortality Data. *British Actuarial Journal*, 14: 257–303.

Vang, Z.M., Sigouin, J., Flenon, A., and Gagnon, A. (2017) Are immigrants healthier than native-born Canadians? A systematic review of the healthy immigrant effect in Canada. *Ethnicity and Health*, 22: 209-241.

Villegas, A.M., and Haberman, S. (2014) On the Modeling and Forecasting of Socioeconomic Mortality Differentials: An Application to Deprivation and Mortality in England. *North American Actuarial Journal*, 18: 168-193.

Wen, J., Cairns, A.J.G., and Kleinow, T., (2021) Fitting Multi-Population Mortality Models to Socio-Economic Groups. *Annals of Actuarial Science*, 15: 144-172.

Wen, J., Kleinow, T., and Cairns, A.J.G. (2020) Trends in Canadian Mortality By Pension Level: Evidence From the CPP and QPP. *North American Actuarial Journal*, 24: 533-561

Wen, J., Cairns, A.J.G., and Kleinow, T. (2023) Modelling Socio-Economic Mortality at Neighbourhood Level. To appear in *ASTIN Bulletin*.

Supplementary Appendix and Figures

A Datasets

Data for England are available at the level of small geographical areas known as Lower Layer Super Output Areas (LSOAs). Each area has typically between 1,000 and 3,000 persons, with an average of about 1,600, across all ages.

There are 32,844 LSOAs at present. New LSOAs are created from time to time in response to growth in housing. Data relating to the LSOAs can be found on the ONS (Office for National Statistics) website www.ons.gov.uk.

A.1 Deaths and exposures

For each LSOA we have:

- data from 2001-2018:
  - death counts, $D(g, i, t, a)$, where g is the sex, i is the LSOA, t is the year, and a is the age last birthday;[12] [13]
  - central exposed to risk, $E(g, i, t, a)$, equated to the mid-year population estimates for 2001-2018 available from the ONS;[14]
- One off (i.e. not observed through time) *predictive variables* for each LSOA that might be associated with higher or lower than average rates of mortality.

A.2 Potential predictive variables and related data

This appendix lists all data considered, with predictive variables used in the final analysis picked out in bold.

1. LSOA index.
   - LSOA codes are of the form "E010xxxxx" where the LSOA index *xxxxx* ranges from 00001 to 33768.

[12] User-requested deaths data used in this study can be found at `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/007807deathsbylowersuperoutputareaageandsexenglandandwales2001to2016`

[13] Death counts are *registrations* in calendar year t rather than *occurrences*. The advantage to the ONS and users of using registrations is that the tables can be produced in a much more timely manner. Death counts by year of occurrence can be delayed by the very small number of deaths that require an inquest.

[14] Mid-year population estimates can be found at `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates`. Note that data for 2012-2016 have been revised slightly since our original download by the ONS to account for revisions made to local authority population estimates (ONS, 2018). Unrevised files for 2015 and 2016 are available on the same web page. The authors have verified that use of the revised data make very little difference to estimates of relative risk, $F_1(i)$ and $F_2(i)$.

- 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. The Index of Multiple Deprivation 2015 (IMD)[15]

   This is the 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.

3. Income deprivation (a domain of the 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;
   - sub-domains include **income deprivation affecting older people**, which measures income deprivation amongst people aged 60 and older.

4. **Employment deprivation** (a domain of the IMD)

   - this measures the proportion of the *working* population in each LSOA who are unemployed.

5. Living environment deprivation (a domain of the IMD)

   - this measures the quality of the living environment (indoors and outdoors);
   - indoors: housing in poor condition; **houses with no central heating**;
   - outdoors: **poor air quality**; road traffic accidents.

6. Barriers to housing and services (a domain of the IMD)

   - this is subdivided under two main headings: 'wider barriers' and 'geographical barriers';
   - wider barriers includes: **overcrowding in households**, **affordability of housing**, and homelessness;

[15] For further details, see Office for National Statistics (2015).

- geographical barriers measures distance to key services. A higher value for geographical barriers implies more 'deprived', but it can also be associated with lower mortality: for example, greater distances to services might indicate that the LSOA is more affluent or rural with housing more spaced out, and, indeed, the geographical barriers variable is negatively correlated with income deprivation.

7. **Average number of bedrooms** (based on 2011 Census data)

   - this measures the average number of bedrooms per household in the LSOA
   - in contrast to the deprivation indices, a high value (more bedrooms) is likely to be associated with lower mortality.

8. Highest level of qualification (data from the ONS)

   - this gives the proportion of a particular group within the LSOA who have attained a particular level of education
   - 6 age groups: All; 16 to 24; 25 to 34; 35 to 49; 50 to 64; 65 plus;
   - 8 education groups:
     (a) All categories: Highest level of qualification
     (b) No qualifications
     (c) Level 1 qualifications
     (d) Level 2 qualifications
     (e) Apprenticeship
     (f) Level 3 qualifications
     (g) Level 4 qualifications and above
     (h) Other qualifications
   - see

`www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels`;

   - In our analysis we used the **proportion of individuals aged 65 and over with level 4 qualifications or higher**. Level 4 and higher can be interpreted as meaning any qualifications that take individuals beyond A-levels/high-school-graduation or equivalent.

9. **Proportion of population born in the UK** (data from the ONS/NOMIS)

   This measures the proportion of usual residents that were born within the UK in an LSOA, calculated by dividing the number of UK-born residents by the total number of residents.

10. **Urban-Rural Classification**

    1 Conurbation: non London

    2 City or town

    3 Rural town

    4 Rural village and dispersed

5 Conurbation: London

11. 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

12. Communal establishments

    - 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 sex;
      - 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.
      - We use the **proportion of the 60+ age group who are in a care home with nursing** and the **proportion of the 60+ age group who are in a care home without nursing**.

B Age Standardised Mortality Rate (ASMR)

B.1 Basic definition

The purpose of the ASMR is to facilitate comparison of mortality rates in different populations. In particular, if the age profiles of different populations are different then some measures (e.g. deaths per 100,000 population) might simply reflect differences in the age profile even if death rates at individual ages are identical. The ASMR avoids this by using a standard population rather than the actual age profile.

The basic definition, ignoring other indices, over the age range (a_0, a_1) is

$$ASMR(a_0, a_1) = \frac{\sum_{a=a_0}^{a_1} m(a)ES(a)}{\sum_{a=a_0}^{a_1} ES(a)}$$

where $ES(a)$ is the standard population at age a[16], and $m(a)$ is the death rate at age a.

B.2 Further development

In this paper we make use of a number of variants of the ASMR. Various age ranges are considered: e.g. 40-49, 40-64, 65-89, etc. We also calculate ASMR's by region and by deprivation or other deciles.

Suppressing indices for the age range, (a_0, a_1), calendar year, t, and sex, g, we start with the following death rates:

- $m(a)$ = national death rate at age a,
- $m_I(i, a)$ = income-deprivation decile i death rate,
- $m_R(r, a)$ = region r death rate,
- $m_{RI}(r, i, a)$ = death rate at age a in region r, deprivation decile i.

Each of these has corresponding ASMRs:

$$\begin{aligned}
ASMR \equiv ASMR(t) &= \frac{\sum_{a=a_0}^{a_1} m(a)ES(a)}{\sum_{a=a_0}^{a_1} ES(a)} \\
ASMR_I(i) &= \frac{\sum_{a=a_0}^{a_1} m_I(i,a)ES(a)}{\sum_{a=a_0}^{a_1} ES(a)} \\
ASMR_R(r) &= \frac{\sum_{a=a_0}^{a_1} m_R(r,a)ES(a)}{\sum_{a=a_0}^{a_1} ES(a)} \\
ASMR_{RI}(r,i) &= \frac{\sum_{a=a_0}^{a_1} m_{RI}(r,i,a)ES(a)}{\sum_{a=a_0}^{a_1} ES(a)}
\end{aligned}$$

For all-cause mortality ($c = 0$) we will be summing over single ages. For cause-of-death mortality ($c > 0$), we will be summing over 5-year age groups: for example if the stated age range is 40 to 49, then we are summing over two age groups: 40-44 and 45-49.

B.3 The Age and Deprivation Standardised Mortality Rate (ADSMR)

Now, we can develop the formula for $ASMR_R(r)$ as follows:

$$ASMR_R(r) = \frac{\sum_{a=a_0}^{a_1} ES(a) \sum_{i=1}^{10} m_{RI}(r,i,a) w_{RI}(r,i,a)}{\sum_{a=a_0}^{a_1} ES(a)} \tag{4}$$

[16]Here we use the European Standard Population, 2013 (Eurostat, 2013). Comparative results in this paper for different populations are unlikely to be sensitive to the choice of standard population.

where $w_{RI}(r,i,a) = E_{RI}(r,i,a)/\sum_j E_{RI}(r,j,a)$ represents the weight carried by decile i amongst deciles 1 to 10, in region r at age a (so, for each r and a, $\sum_i w_{RI}(r,i,a) = 1$).

We then note than some regions have greater proportions of more deprived areas than other regions. The resulting differences in weights then means that some regions will have naturally higher ASMRs even if there are no differences in death rates at the level of income deprivation between regions (i.e. $m_{RI}(r,i,a) = m_I(i,a)$ for all $r = 1,\ldots,9$).

To remedy this, we propose the ADSMR as an alternative to the regional $ASMR_R$s. Specifically we replace the weights $w_{RI}(r,i,a)$ in (4) by $\tilde{w}_{RI}(r,i,a) = 0.1$. Hence

$$\begin{aligned} ADSMR(r) &= \frac{\sum_{a=a_0}^{a_1}\sum_{i=1}^{10} m_{RI}(r,i,a)\tilde{w}_{RI}(r,i,a)ES(a)}{\sum_{a=a_0}^{a_1} ES(a)} \\ &= \frac{\frac{1}{10}\sum_{i=1}^{10}\sum_{a=a_0}^{a_1} m_{RI}(r,i,a)ES(a)}{\sum_{a=a_0}^{a_1} ES(a)} \\ &= \frac{1}{10}\sum_{i=1}^{10} ASMR_{RI}(r,i). \end{aligned}$$

The use of the ADSMR allows us to filter out the impact of differences in deprivation levels. Any differences that remain need further investigation.

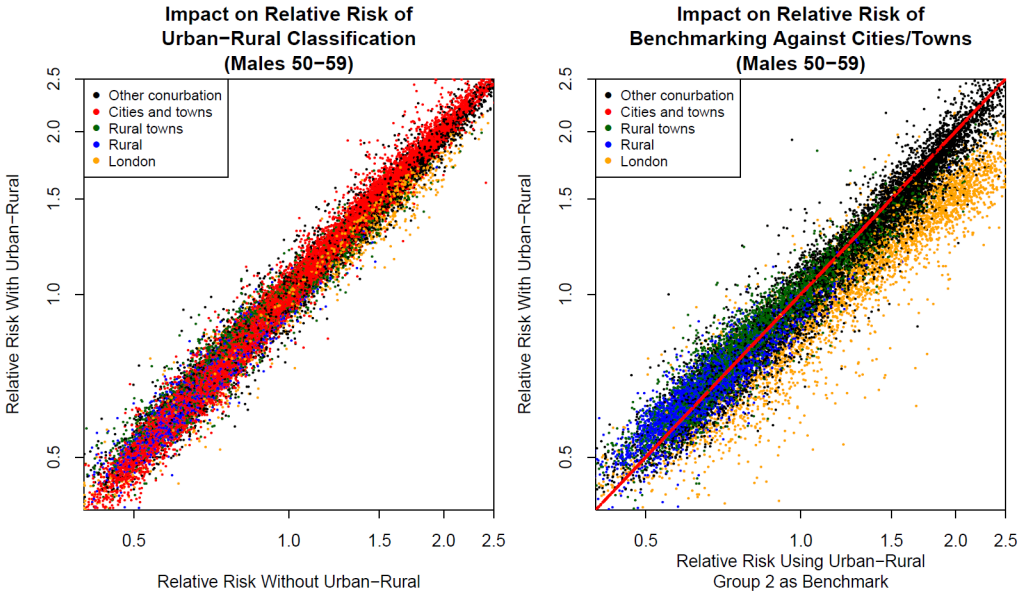

Figure 11: Left: Comparison of socio-economic relative risk, $\hat{F}_1(i)$, in the two cases D (where urban-rural class is not used as a predictive variable) versus A (where each LSOA is benchmarked against its own urban-rural class). Scatterplot of $(\hat{F}_1^D(i), \hat{F}_1^A(i))$. Right: Comparison of socio-economic relative risk, $\hat{F}_1(i)$, in the two cases E (urban-rural class 2 is used as a benchmark) versus A (where each LSOA is benchmarked against its own urban-rural class). Scatterplot of $(\hat{F}_1^E(i), \hat{F}_1^A(i))$.

C Impact of urban-rural class

Here we compare the baseline results with two cases:

- Case D: urban-rural class plays no role in the local linear regression. This produces relative risks $\hat{F}_1^D(i)$ as well as $\hat{F}_2^D(i)$.
- Case E: all LSOAs benchmarked against urban-rural class 2 (cities and towns).[17] This produces relative risks $\hat{F}_1^E(i)$ as well as $\hat{F}_2^E(i)$.

Unlike cases B and C, each of cases D and E require the model to be refitted.

In case D, we simply treat each LSOA as though no information about the urban-rural class is available or this characteristic is not considered relevant, and therefore excluded from modelling. In Figure 11 we show a scatterplot of $\hat{F}_1^D(i)$ versus $\hat{F}_1^A(i)$ coloured by urban-rural class for ages 50-59. If there was no urban-rural effect then all points would lie on the $x = y$ diagonal. In reality, for example, the blue dots (UR class 4: very rural) are less steep indicating that this class is less sensitive to changes in the socio-economic predictive variables than other urban-rural classes. Rural towns (green) and London (orange) can also be seen to be slightly less steep.

When we consider the combined relative risk, $\hat{F}_1(i)\hat{F}_2(i)$, we find that cases A and D produce similar results in terms of goodness of fit in each age group. However, the

[17] Class 2 is the largest of the five classes and has the widest range of predictive variables, allowing LSOAs in other classes to be sensibly matched with LSOAs with similar predictive variables in class 2.

influence of $\hat{F}_2(i)$ is quite different. Specifically, the variance of $\hat{F}_2^D(i)$ is significantly higher than $\hat{F}_2^A(i)$. This goes against our objective to minimise variation in the remaining spatial relative risk in order to explain as much of the observed variation as possible using socio-economic and non-geographical predictive variables.

In case E, we benchmark all LSOA's against their nearest neighbours socio-economically, in urban-rural class 2. Thus we use weights derived from the modified distances:

$$d_1(i,j) \;=\; \begin{cases} ||X(i) - X(j)||_S & \text{if } u(j) = 2 \\ \infty & \text{if } u(j) \neq 2 \end{cases}$$

that is, we only assign a non-zero weight to LSOAs in urban-rural class 2. Results are plotted in Figure 11. Here, we plot $\hat{F}_1^E(i)$ versus $\hat{F}_1^A(i)$. Red dots for class 2 lie on a straight line as there is no change in how class 2 is fitted. Urban-rural classes 4 (rural, blue dots) and 5 (London, orange dots) stand out as being well below the the main diagonal. This suggests that, on a like for like basis (i.e. similar socio-economic predictive variables) LSOAs in class 2 (cities and towns) have significantly higher mortality than very rural areas and in London, particularly at the more deprived end of the spectrum. Systematic differences by urban-rural class highlighted here diminish with age.

D Analysis of residuals

Local linear regression itself does not require any assumption about the distribution of the deaths, simply that $E[D(i)] = \hat{D}_0(i)F_1(i)F_2(i)$. In our analysis of the residuals we refine this further by investigating if the death counts have a Poisson distribution.

We propose here the use of *randomised probability-transformed residuals*. These are calculated as follows:

- Suppose we have observations $w_1, \ldots, w_n$ of the random variables $W_1, \ldots, W_n$ with the null hypothesis that $W_i \sim \text{Poisson}(\lambda_i)$ and that the W_i are independent of one another.
- For each $i = 1, \ldots, n$, let

$$\begin{aligned} q_{0i} &= Pr(W_i < w_i|\lambda_i) \\ q_{1i} &= Pr(W_i \leq w_i|\lambda_i) \end{aligned}$$

  where the probabilities are calculated under the Poisson assumption with mean λ_i. If $w_i = 0$ then $q_{0i} = 0$.
- Simulate the *randomised probability-transformed residual* $U_i \sim U(q_{0i}, q_{1i})$.[18]

If the Poisson null hypothesis is true then the U_i will be independent and identically distributed uniform$(0,1)$ random variables. In our mortality setting, each U_i is derived from the observed deaths, $D(i)$, and the estimated mean, $\hat{D}_0(i)F_1(i)F_2(i)$. This provides us with the potential for a number of graphical diagnostics. These could include, for example, QQ plots and histograms of the U_i. Here we consider two types of scatterplots.

[18] That is, uniformly distributed on the interval (q_{0i}, q_{1i}).

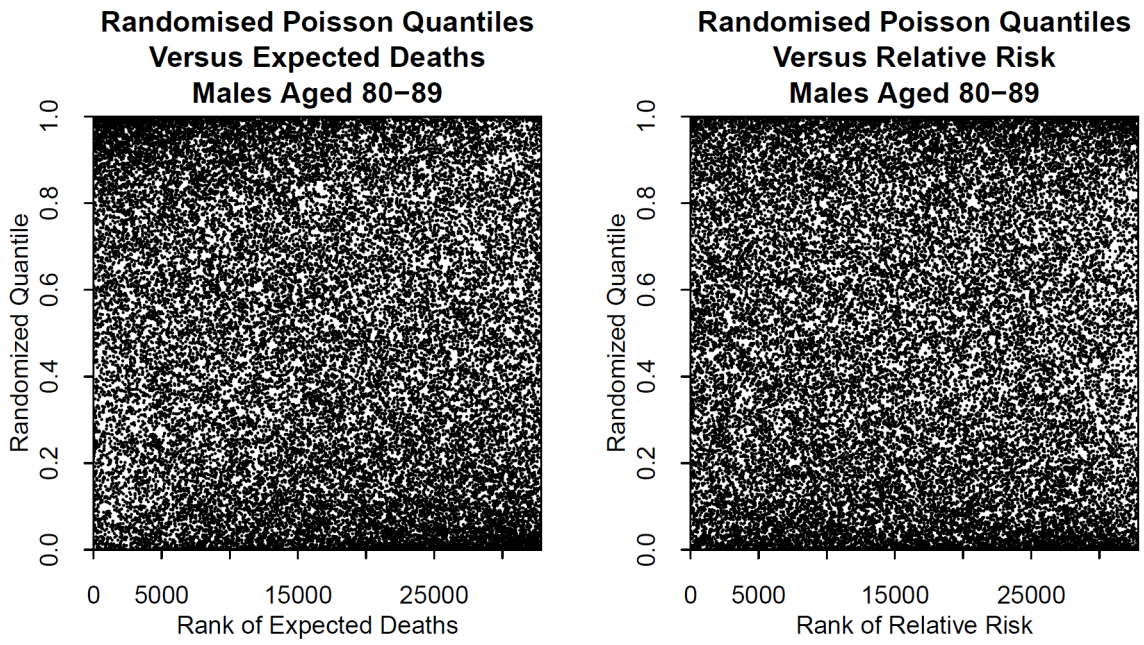

Figure 12: Scatterplots of the ranks of the expected deaths (left) and the relative risk, $\hat{F}_1(i)\hat{F}_2(i)$, (right) versus the randomised probability-transformed residual, U_i, for males aged 80-89.

- Plot the (V_i, U_i) where V_i is the rank of the expected deaths $\hat{D}_2(i) = \hat{D}_0(i)\hat{F}_1(i)\hat{F}_2(i)$ out of $\hat{D}_2(1), \ldots, \hat{D}_2(L)$. This is illustrated in the left-hand panel of Figure 12 males aged 80-89.
- Plot the (W_i, U_i) where W_i is the rank of the combined relative risk $\hat{F}(i) = \hat{F}_1(i)\hat{F}_2(i)$ out of $\hat{F}(1), \ldots, \hat{F}(L)$. This is illustrated in the right-hand panel of Figure 12.

In both cases, if the Poisson hypothesis is true then the scatterplots should look uniform and random with no clustering[19].

For younger age groups the scatterplots do look reasonably uniform reasonably uniform and random. But for ages 80-89 both scatterplots in Figure 12 are less uniform indicating that there are further effects that have not been captured in the model.

In the left-hand panel we can see some clustering in the top left and bottom right. In the right-hand panel the clustering is more evenly distributed along the top and bottom for ages 80-89: an indication of overdisperion of some sort. The differences in clustering in the two panels are consistent with the fact that exposures are *estimated* exposures and, therefore, subject to estimation error. The pattern in the left-hand panel is simply telling us that if exposures are underestimated then expected deaths are low (so towards the left of the scatterplot) and deaths tends to be higher than estimated (so a high value of U_i). Similarly, if the exposures are overestimated then expected deaths are high (so towards the right of the scatterplot) and deaths tend to be lower than estimated (so a low value of U_i). This illustrates just one impact of the challenge that the ONS faces when estimating

[19]Note that the V_i and W_i are evenly distributed on the integers $1, \ldots, L$ rather than randomly distributed.

the population between censuses in each LSOA. Figure 12 suggests that the problem is only significant at higher ages.[20]

E Supplementary plots

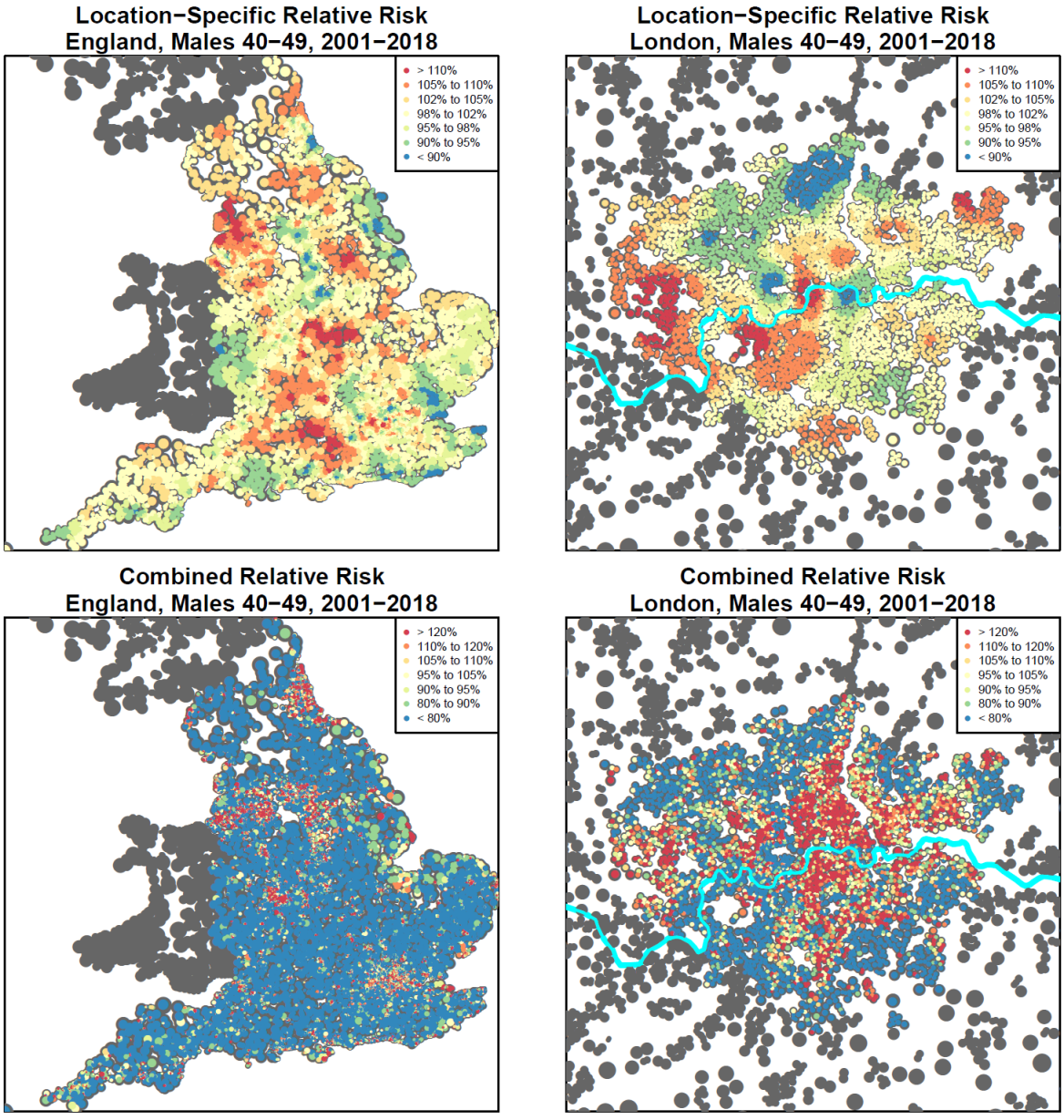

Figure 13: Top row: estimated spatial relative risk, $F_2(i)$, by LSOA for England (left) and London (right) for males, ages 40-49. Bottom row: combined relative risk, $F_1(i)F_2(i)$. Dot sizes reflect the physical size of each LSOA.

[20]Note, the the presence of overdispersion has no direct impact on the estimation methodology, which only relies on the expected number of deaths rather than the distribution around that mean. Methods such as Generalised Linear Models would, on the other hand, would need to specify how to handle overdispersion: e.g. by replacing the Poisson assumption above with the negative binomial distribution.

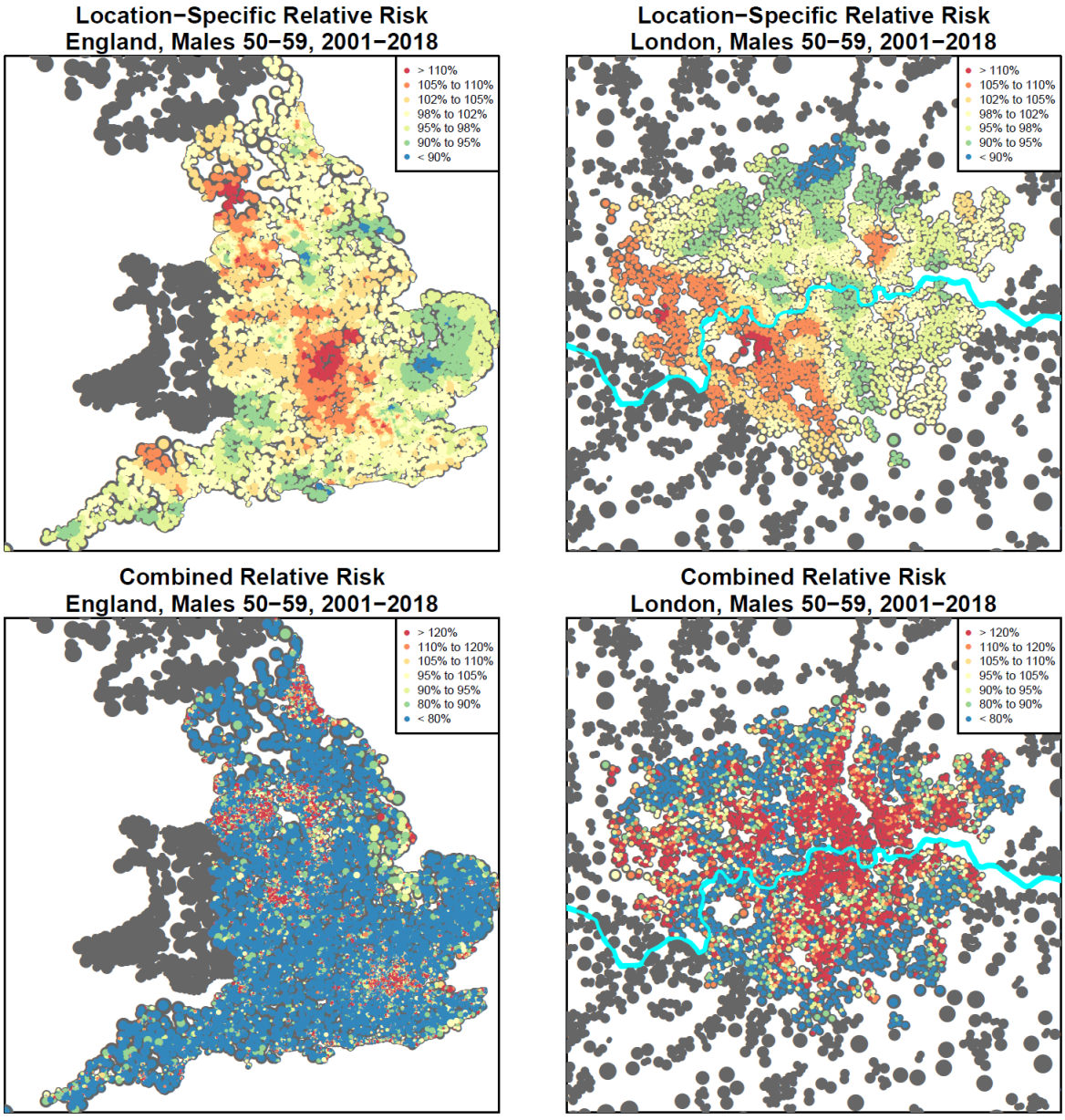

Figure 14: As Figure 13 but for ages 50-59.

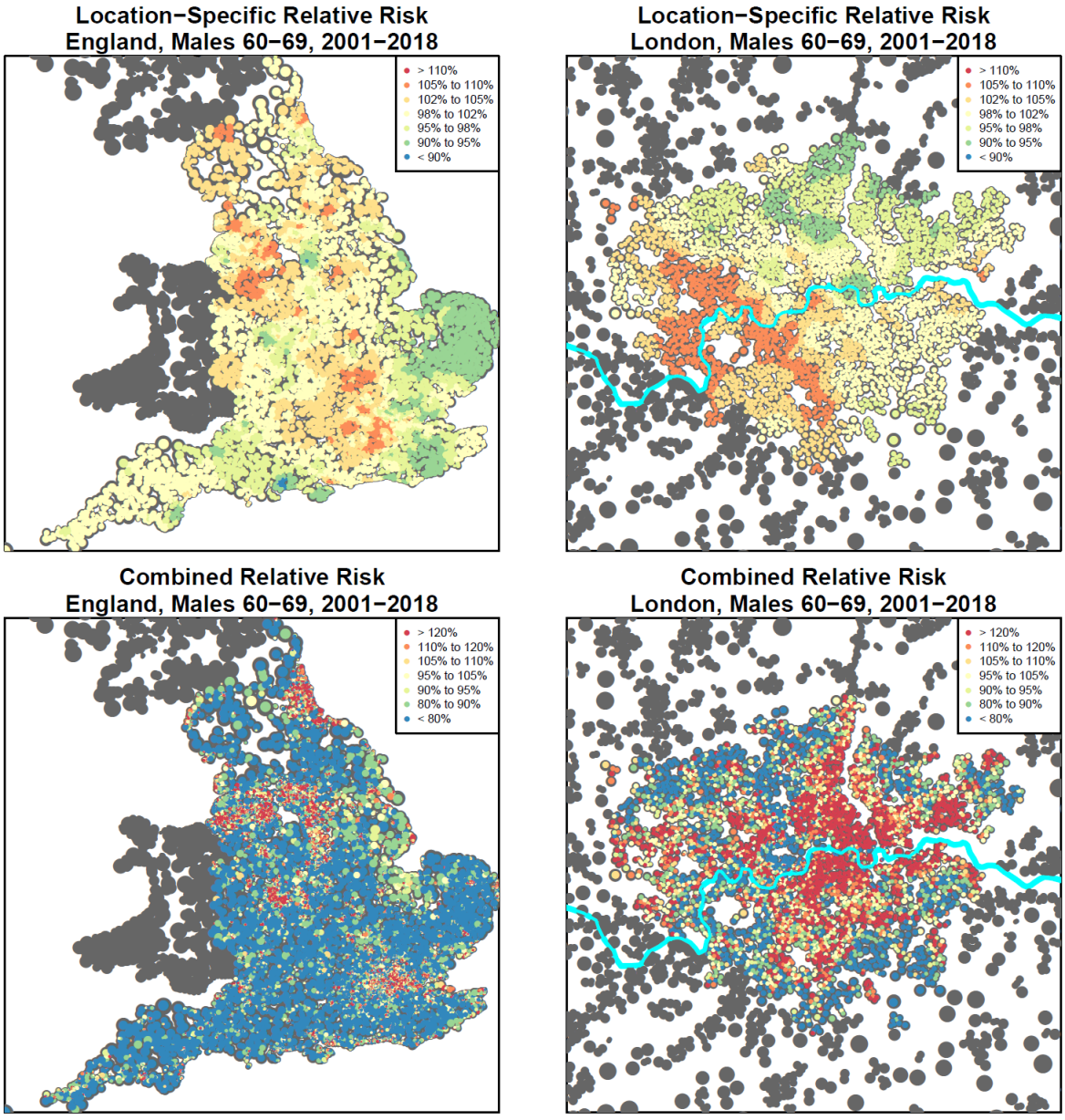

Figure 15: As Figure 13 but for ages 60-69.

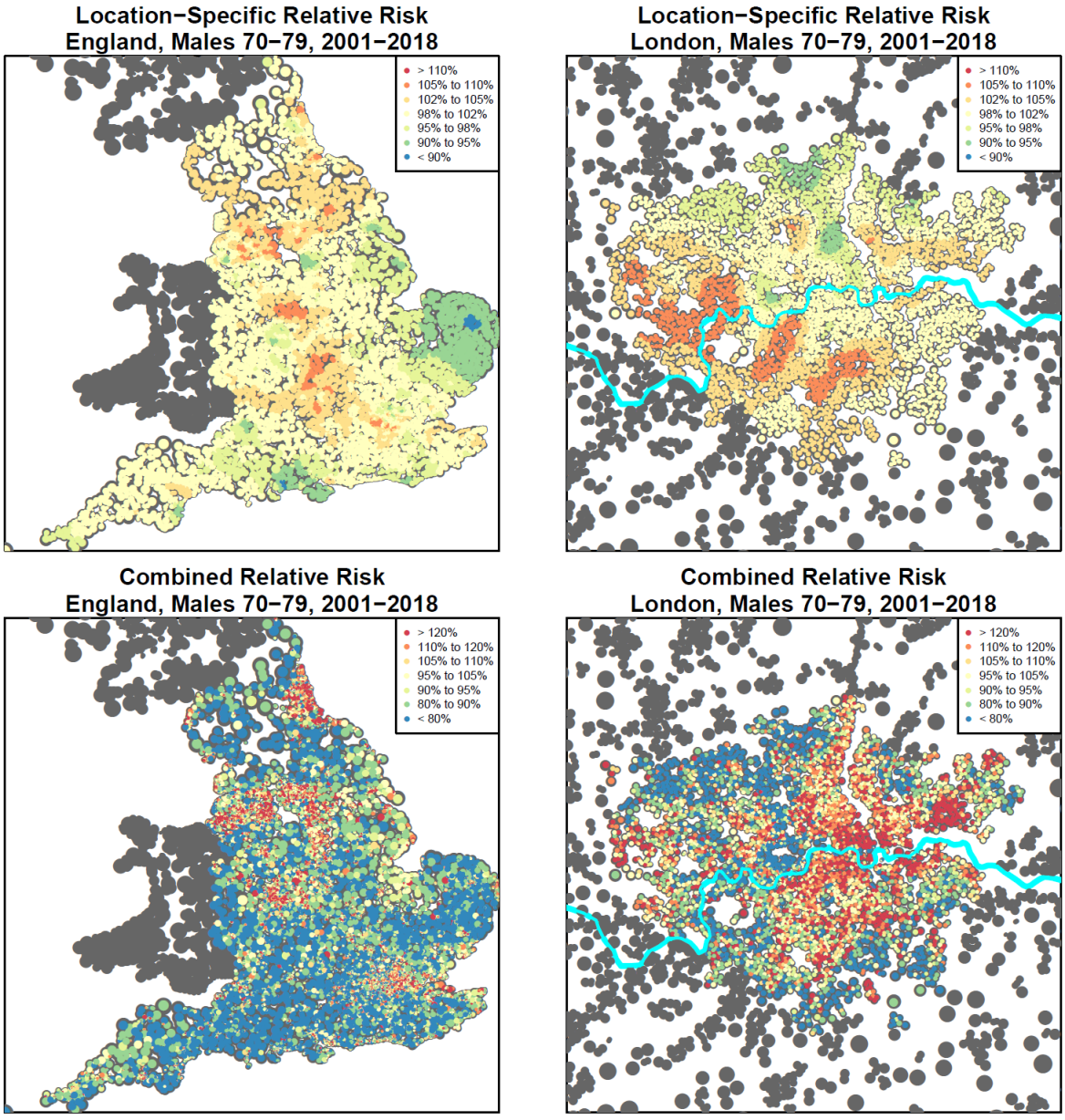

Figure 16: As Figure 13 but for ages 70-79.

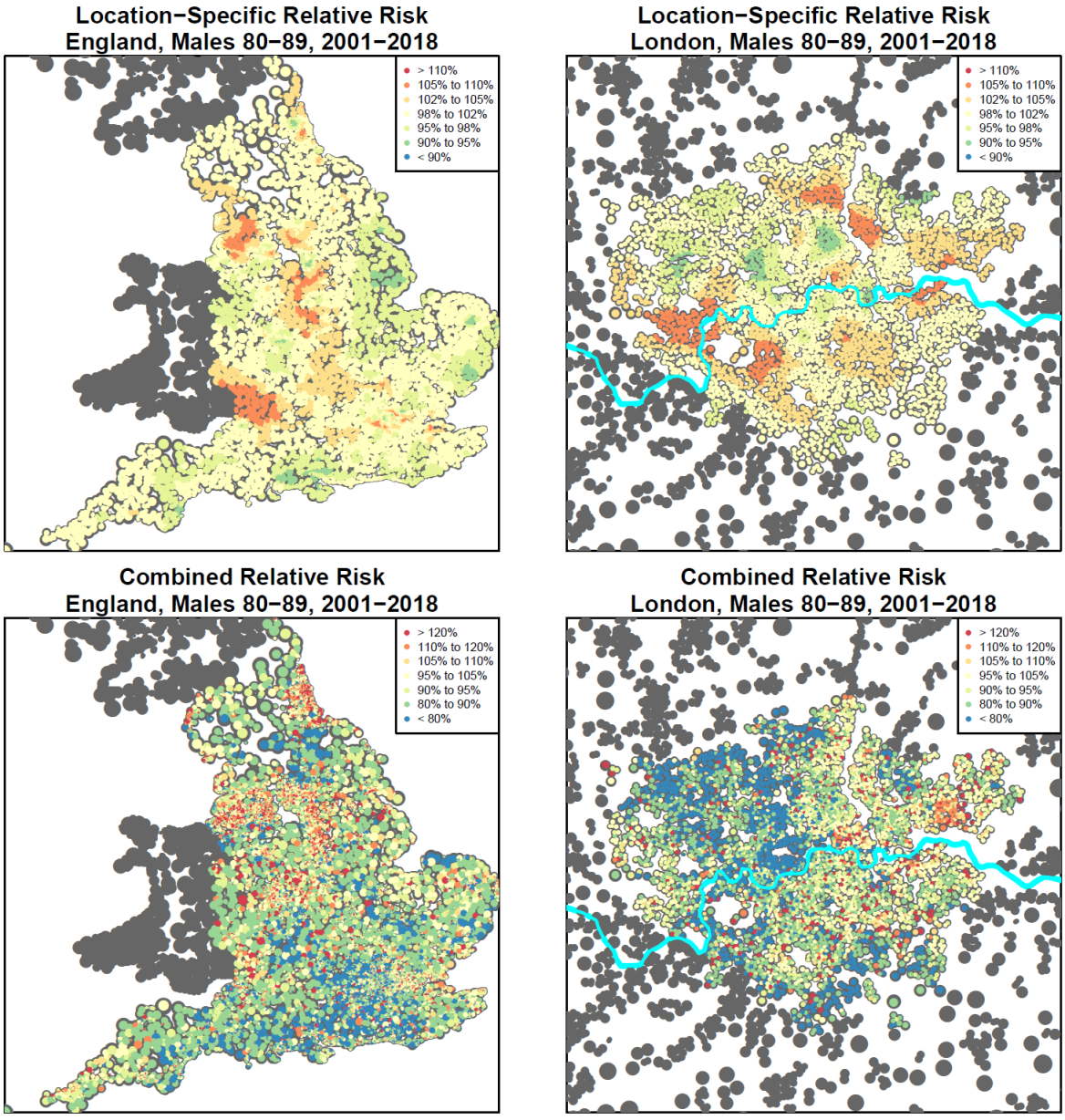

Figure 17: As Figure 13 but for ages 80-89.

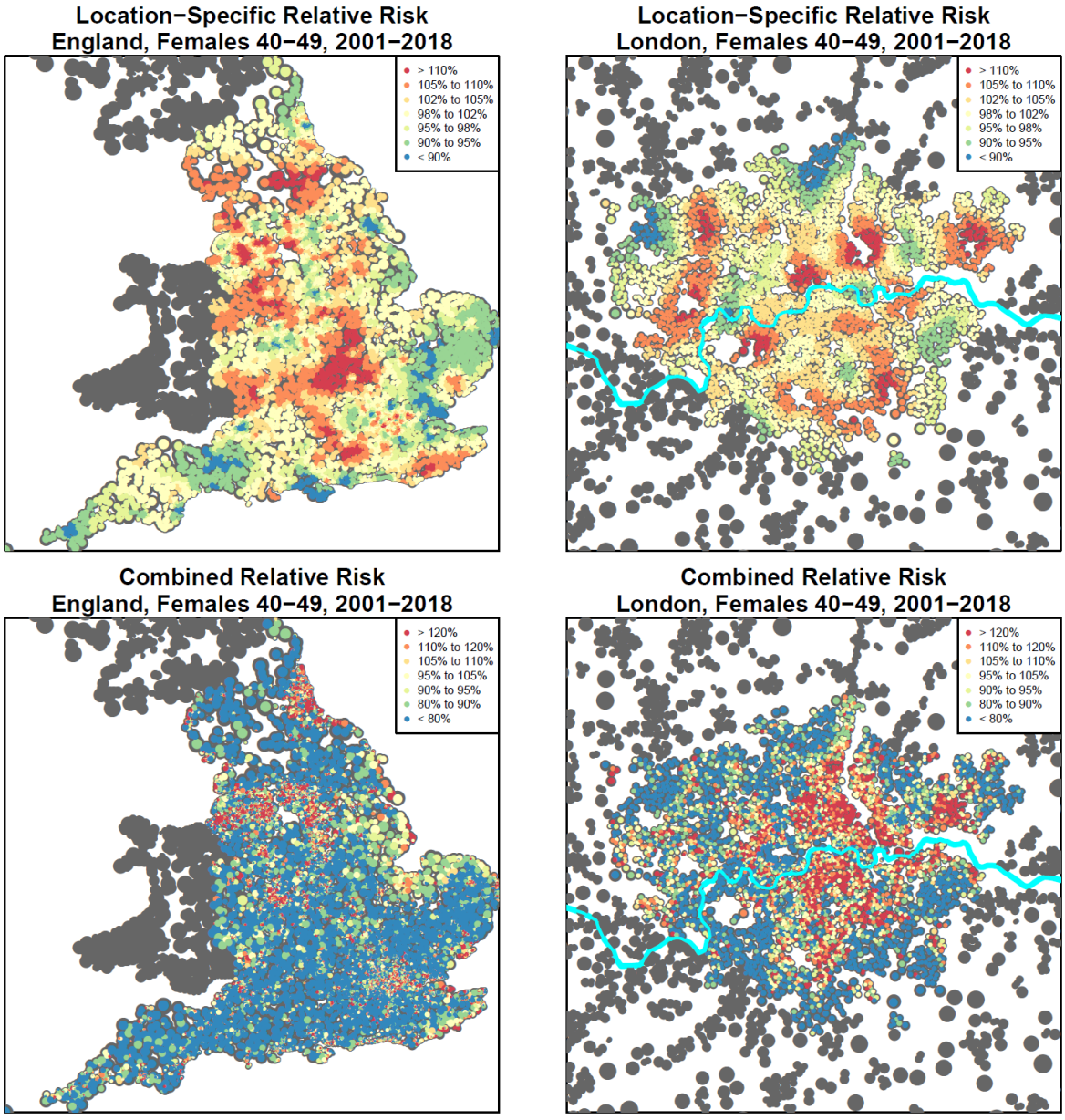

Figure 18: Top row: estimated spatial relative risk, $F_2(i)$, by LSOA for England (left) and London (right) for females, ages 40-49. Bottom row: combined relative risk, $F_1(i)F_2(i)$. Dot sizes reflect the physical size of each LSOA.

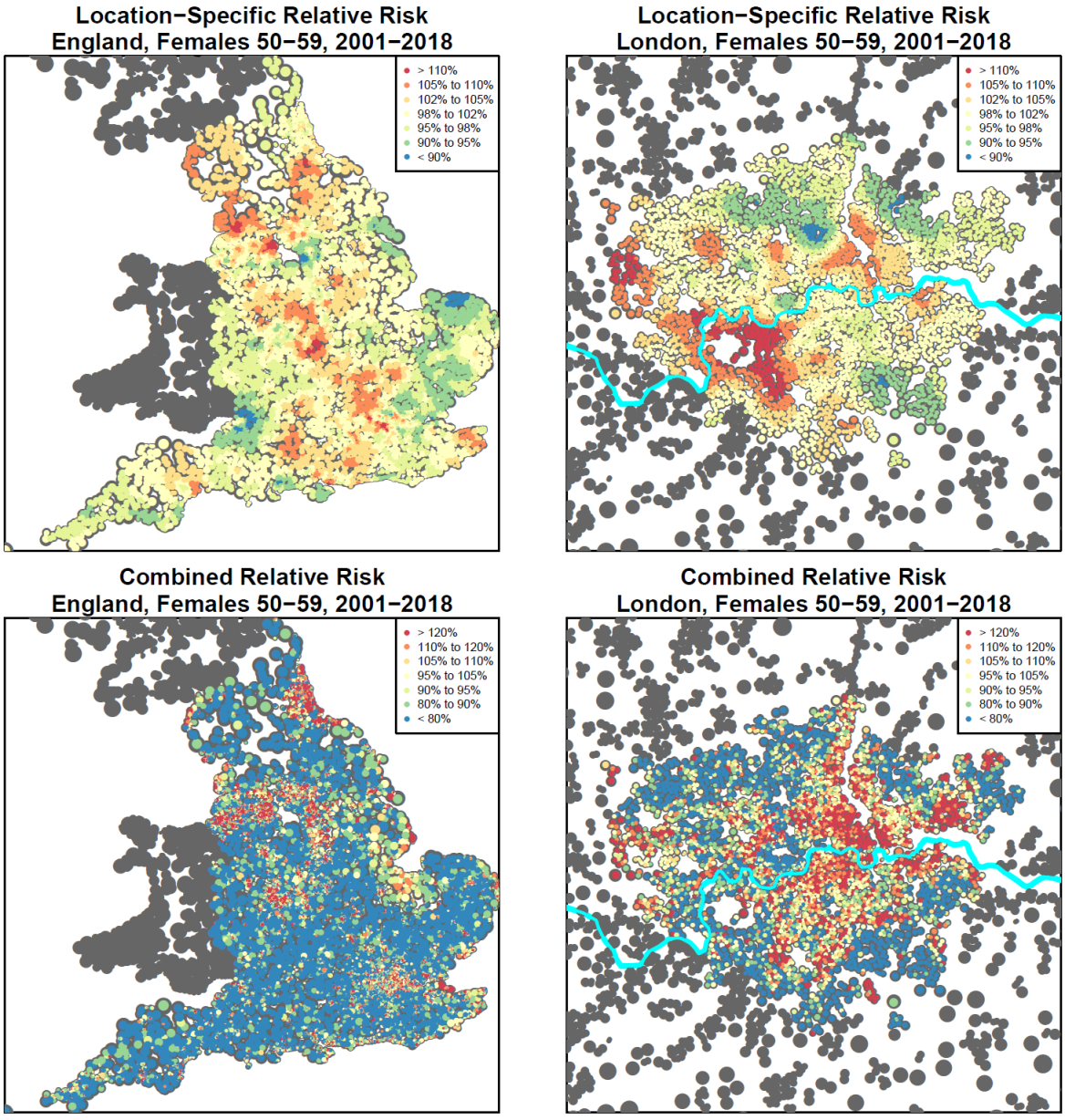

Figure 19: As Figure 18 but for ages 50-59.

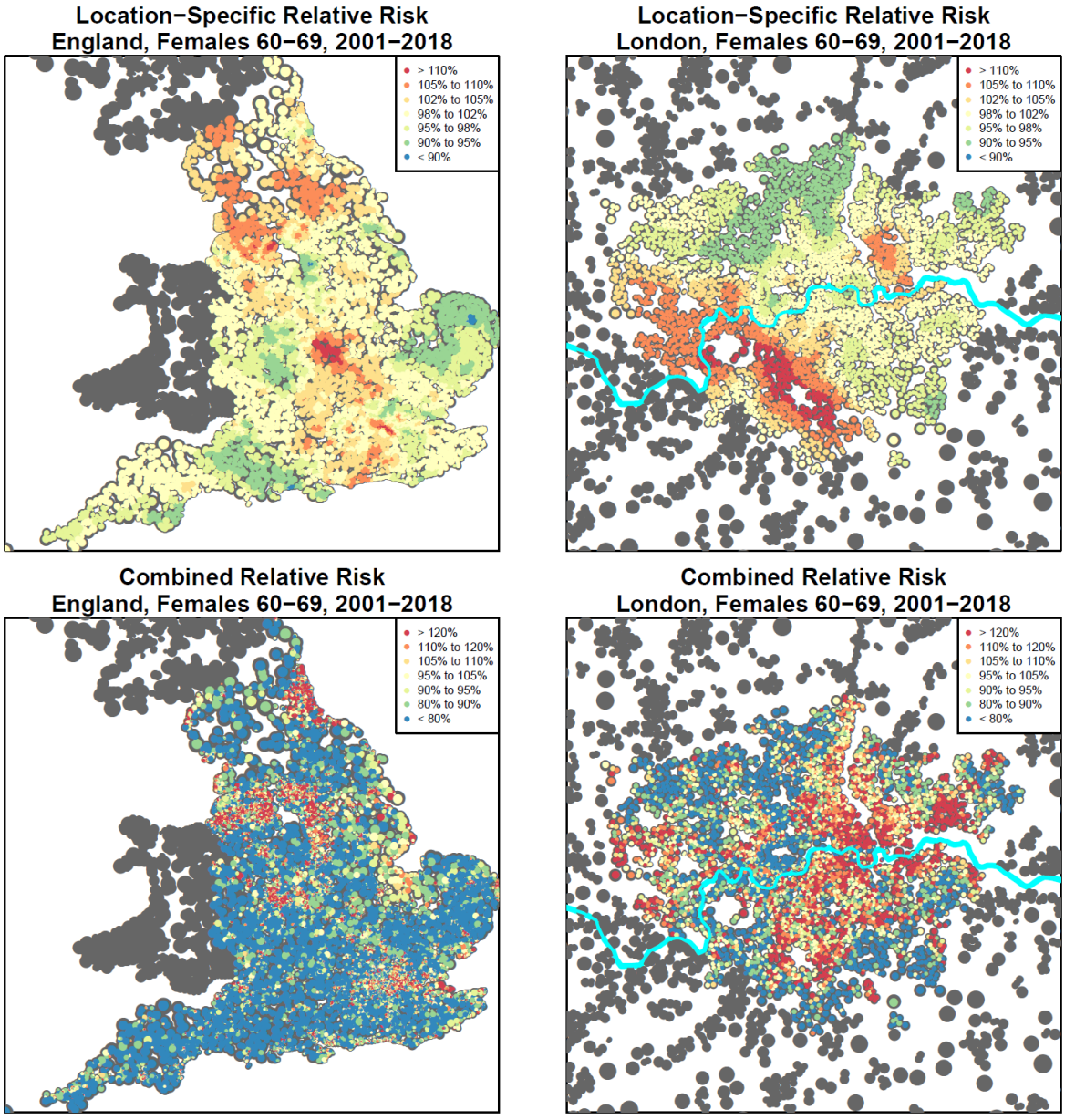

Figure 20: As Figure 18 but for ages 60-69.

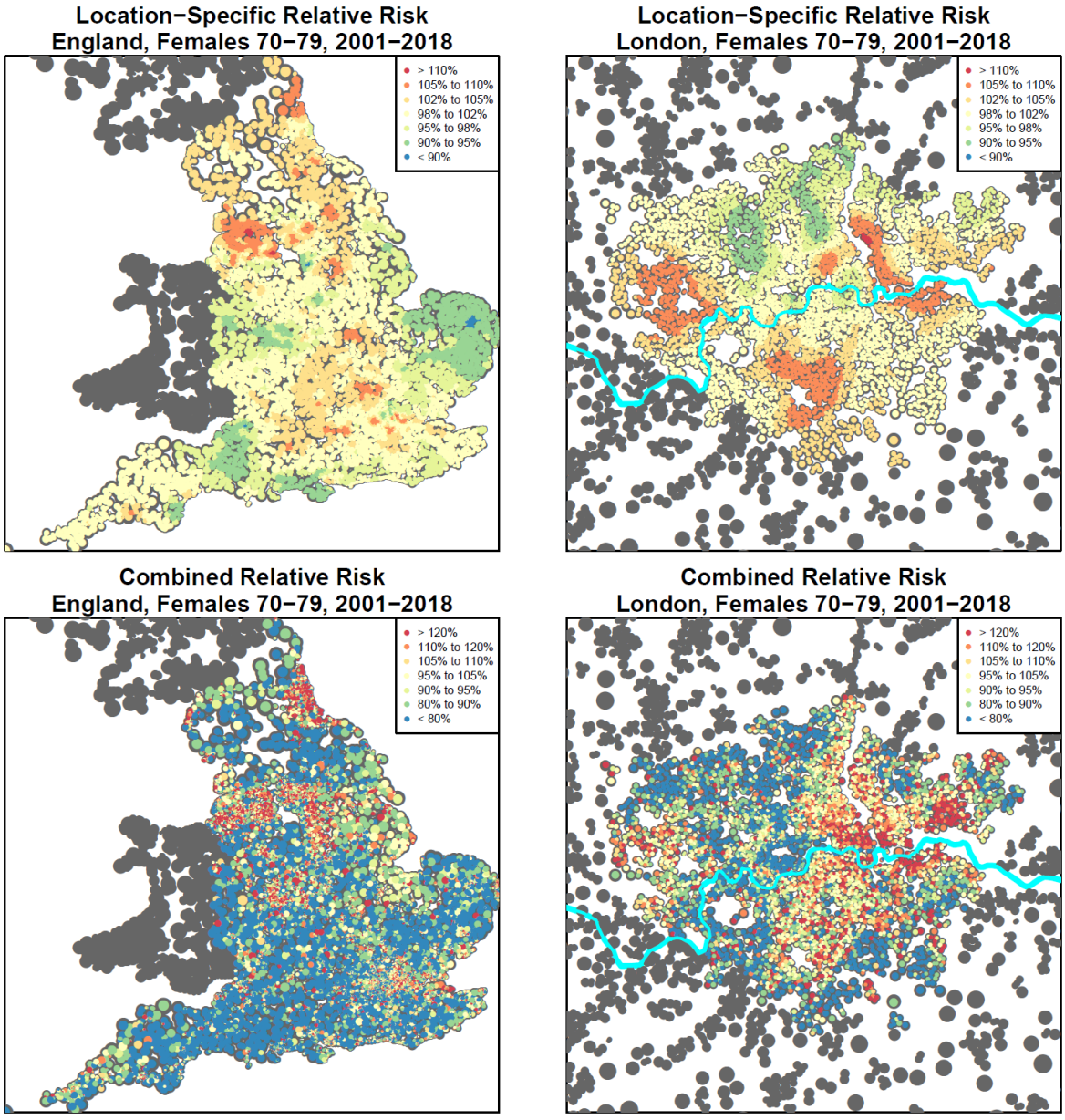

Figure 21: As Figure 18 but for ages 70-79.

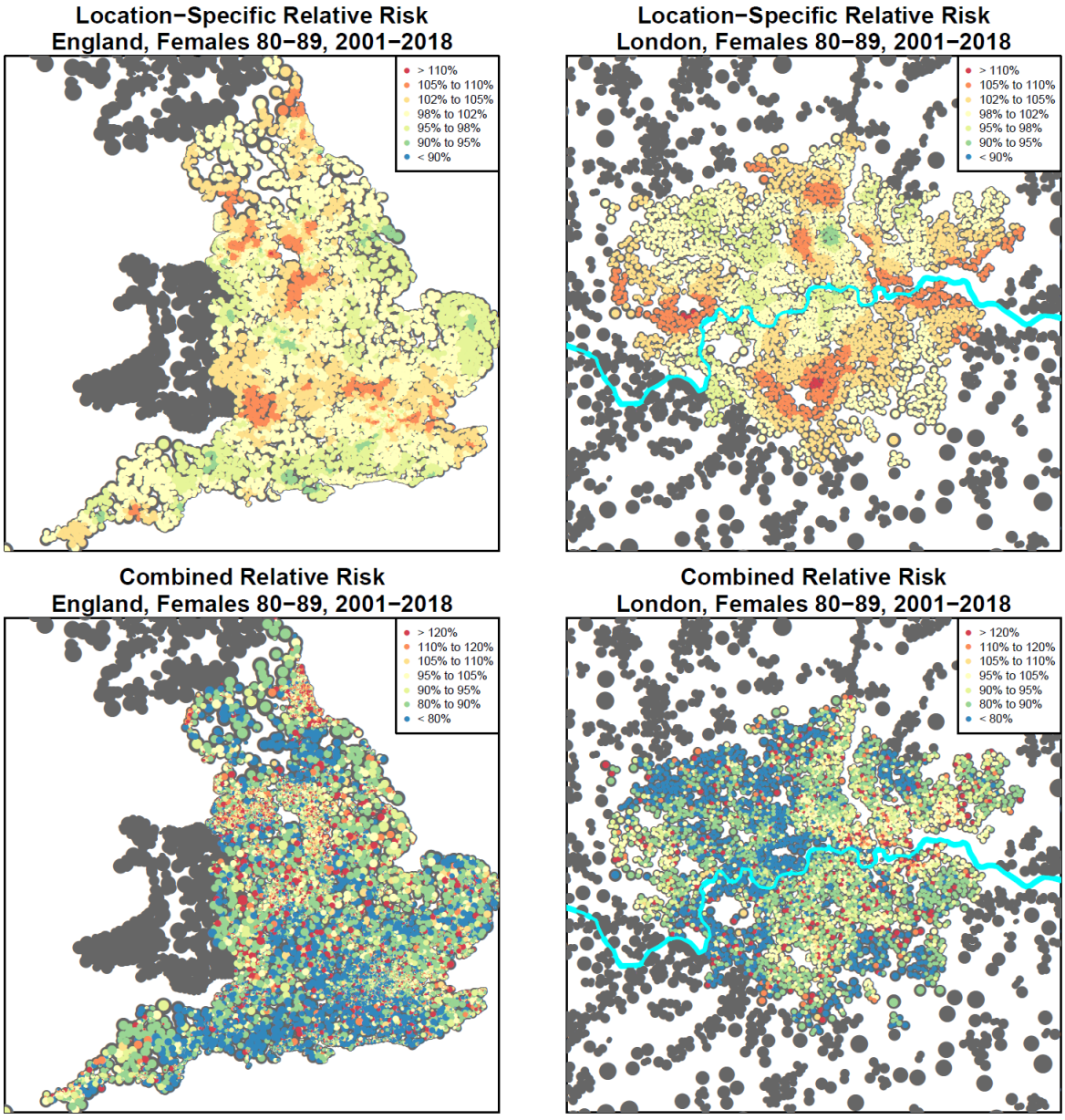

Figure 22: As Figure 18 but for ages 80-89.

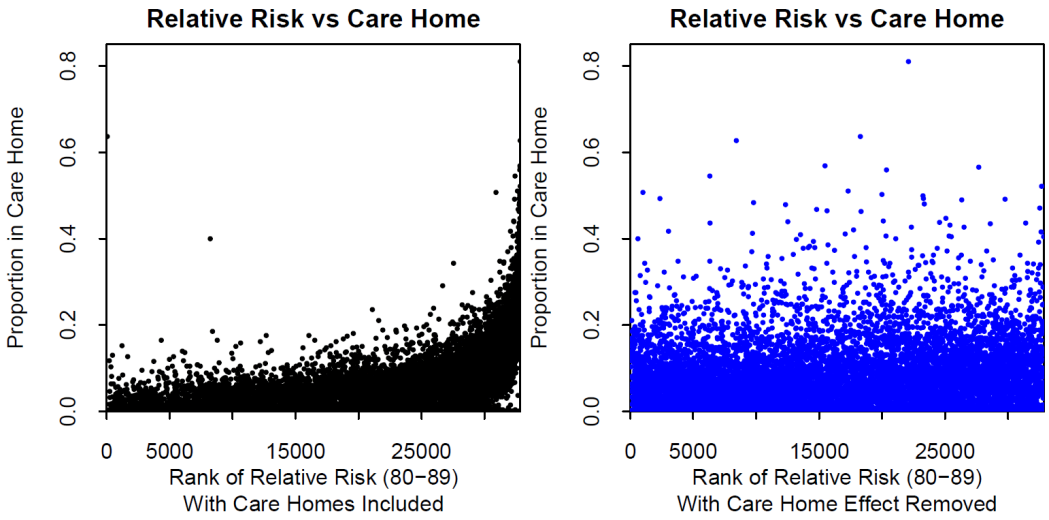

Figure 23: Left: scatterplot of the rank of $\hat{F}_1^A(i)\hat{F}_2(i)$ for the 80-89 age group versus $X_7(i) + X_8(i)$. Right: scatterplot of the rank of the relative risk of the non-care-home population relative risk $\hat{F}_1^B(i)\hat{F}_2(i)$ for the 80-89 age group versus $X_7(i) + X_8(i)$.