Astin Bulletin (2023), **53**, pp. 285–310
doi:10.1017/asb.2023.12

International Actuarial Association
Association Actuarielle Internationale

RESEARCH ARTICLE

Modelling socio-economic mortality at neighbourhood level

Jie Wen[1], Andrew J.G. Cairns[2,3] and Torsten Kleinow[4,*]

[1]Lloyds Banking Group, Edinburgh EH3 9PE, UK, [2]The Maxwell Institute for Mathematical Sciences, Edinburgh EH9 3FD, UK, [3]Department of Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK and [4]Research Centre for Longevity Risk, Faculty of Economics and Business, University of Amsterdam, Amsterdam, Netherlands
[*]Corresponding author. E-mail: t.kleinow@uva.nl

Received: 30 September 2022; **Revised:** 7 March 2023; **Accepted:** 7 March 2023; **First published online:** 11 April 2023

Keywords: Socio-economics; random forest; mortality; longevity; lower layer super output area

Abstract

In this study, we quantify the relationship between socio-economic status and life expectancy and identify combinations of socio-economic variables that are particularly useful for explaining mortality differences between neighbourhoods in England. We achieve this by examining socio-economic variation in mortality experiences across small areas in England known as lower layer super output areas (LSOAs). We then consider 12 socio-economic variables that are known to have a strong association with mortality. We estimate the relationship between those variables and mortality rates using a random forest algorithm. Based on the resulting estimate, we then create a new socio-economic mortality index – the Longevity Index for England (LIFE). The index is constructed in a way that eliminates the impact of care homes that might artificially increase mortality rates in LSOAs with care homes compared to LSOAs that do not contain a care home. Using mortality data for different age groups, we make the index age-dependent and investigate the impact of specific socio-economic characteristics on the age-specific mortality risk. We compare the explanatory power of the LIFE index to the English Index of Multiple Deprivation (IMD) as predictors of mortality. While we find that the IMD can explain regional mortality differences to some extent, the LIFE index has significantly greater explanatory power for mortality differences between regions. Our empirical results also indicate that income deprivation amongst the elderly and employment deprivation are the most significant socio-economic factors for explaining mortality variation across LSOAs in England.

1. Introduction

It is well documented that there is a strong association between mortality and socio-economic status. While this relationship has been known for many years, the availability of more granular data allows us to look more closely at the impact of socio-economic characteristics on mortality. In recent years, there have been numerous studies on this relationship: Bennett *et al*. (2015) discuss modelling life expectancies in different areas of England and Wales via a Bayesian model with spatial effects; Raleigh and Kiri (1997) study trends in life expectancy in relation to deprivation; Woods *et al*. (2005) describe mortality in England and Wales by deprivation and in each government office region during 1998; Cairns *et al*. (2019) identify different socio-economic groups in Denmark and model their mortality rates using an affluence index and Wen *et al*. (2021) explore mortality rates in populations identified by deciles of the English Index of Multiple Deprivation (IMD). Deciles of the IMD index are also used by Lyu *et al*. (2022) to identify three socio-economic groups that are then used to study the effectiveness of an index-based longevity hedge for which the three groups are modelled with a generalised three-way Li–Lee model. Mayhew *et al*. (2020) also use the IMD to measure deprivation and study its impact on demographic differences measured by life expectancy, lifespan variation and mortality. They also consider differences in those quantities between different geographic areas. Finally, the IMD was also used by Villegas and Haberman (2014) who proposed a new approach for the joint modelling of the mortality in subpopulations of a larger national population. All of those studies documented significant differences in mortality levels and mortality improvement rates between socio-economic classes.

In this paper, we develop a socio-economic mortality index that we call the Longevity Index for England – LIFE. The LIFE index is designed to enable us to predict mortality risk in small neighbourhoods across England relative to national mortality based on specific socio-economic data for such neighbourhoods. In other words, the LIFE index models the relative risk of dying within a small population in England as a function of certain socio-economic variables. The index is created based on a regression analysis using a random forest algorithm to estimate a non-parametric regression function.

We consider a total of 12 socio-economic factors that are known to be linked to mortality. Many are domains or subdomains of the English indices of deprivation: in particular, the IMD – English Index of Multiple Deprivation. Those indices are constructed and published by the Ministry of Housing, Communities & Local Government in the United Kingdom. Others are derived from the 2011 national census. We are using the indices of deprivation published in 2015, see Smith *et al*. (2015). The indices of deprivation provide a score for small neighbourhoods, called lower layer super output area (LSOAs), in different domains of deprivation. An LSOA is a lower layer super output area – a small area in England with a population of about 1500 people. For the 2015 indices of deprivation, there were 32,844 LSOAs in England.

The main difference to the IMD-based research mentioned above is that we consider the impact of individual variables on mortality in very small neighbourhoods rather than grouping the data by one measure (IMD deciles) into rather large groups that are then considered to be homogenous.

Two of the 12 variables used to explain mortality in LSOAs are measuring the proportion of an LSOA's population that live in care homes. Since the existence of care homes has the potential to artificially increase the mortality in an LSOA, we adjust those variables for the construction of the LIFE index.

In the empirical part of this paper, we investigate the explanatory power of the LIFE index in comparison to the English Index of Multiple Deprivation. We find that the LIFE index is better able to explain regional differences in mortality than the IMD. This comparison is based on the analysis of age and deprivation standardised mortality rates (ADSMRs). We explain the construction of ADSMRs and show that large mortality differences between regions remain unexplained when ADSMRs are based on the IMD. Applying LIFE-based ADSMRs reduces those regional differences significantly indicating that the LIFE index is better able to capture mortality related factors. We also study the impact of specific predictors on LIFE scores. It turns out that old-age income deprivation is the most powerful predictor of mortality amongst those considered.

As mentioned above the index is based on an application of the random forest algorithm to estimate the non-parametric regression function linking mortality to the 12 predictors, see Breiman (2001) and James *et al*. (2013). This approach offers a high degree of flexibility, does not require assumptions about the underlying relationship between predictive variables, and we show that it is an effective tool for analysing large and complex mortality datasets. The random forest estimator is a well-understood and widely used method from machine learning, and it has been applied in mortality modelling as an alternative to parametric models, see for example, Bjerre (2022), Hong *et al*. (2021), Levantesi and Nigri (2020) and Levantesi and Pizzorusso (2019). While those authors apply the random forest method to improve the goodness-of-fit and the predictive power of mortality models, we use the method to analyse the impact of socio-economic characteristics on higher or lower mortality in any given neighbourhood compared to the national mortality levels for a population with a similar age structure.

Estimating the effect of socio-economic factors on the deviation of LSOA-specific mortality from national mortality is a regression problem. As for many other regression problems, there are several estimation methods to choose from. We have chosen the random forest method as it offers a high degree of flexibility, but in contrast to other non-parametric methods it requires a relatively small number of hyper-parameters to be chosen. In addition, computations are rather fast.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

Wen (2022) applies other estimators to explain the impact of socio-economic factors on mortality. He finds that non-parametric methods (random forest and local linear regression) outperform generalised linear models in terms of an out-of-sample mean squared error. It seems that the structure of the generalised linear models considered by Wen (2022) is not sufficiently flexible to capture non-linear effects of certain factors or the join effects that some factors might have on mortality. It might be possible to extend the analysis by Wen (2022) by including GLMs with more factors or interaction terms between certain factors, but this is not in the scope of this study.

The remainder of this paper is organised as follows. In Section 2, we describe the mortality data and socio-economic data used in our study, and in Section 3, we introduce the LIFE index. Section 4 provides an overview of the random forest method and shows how it is applied in the context of this paper. In this section, we also study the performance of this method for the data in our study. We then investigate the impact of different age groups on the LIFE index ranks of LSOAs in Section 5 and compare the LIFE index to the IMD in Section 6. We then apply the life index to study the distribution of low and high mortality groups across urban and rural LSOAs in Section 7 and analyse the impact of individual variables in Section 8. We return to a comparison of the LIFE index with the IMD in Section 9 where we consider mortality rates in LIFE deciles and IMD deciles and compare ADSMRs based on the two indices. Our final conclusions are presented in Section 10.

2. Data

The data used in this paper are for England and have been sourced from the UK's Office for National Statistics (ONS). Further details can be found in the supplementary material published online and in Wen (2022).

Socio-economic data and mortality data are available at a neighbourhood level called LSOA. An LSOA is a geographical unit that describes a small neighbourhood with a population size of around 1600, and generally with a high degree of socio-economic homogeneity within each LSOA. The number of LSOAs and their boundaries varies from time to time as populations change. This paper uses the revisions based on the 2011 Census and there are $N = 32,844$ LSOAs. The data described below are available for each of the N LSOAs.

The specific data considered in our study are the following

- mid-year population estimates (exposure size) E_{ita} by single LSOA $i = 1, \ldots, N$, year t and age a;
- death counts D_{ita} by single LSOA $i = 1, \ldots, N$, year t and age a;
- a vector of K predictive variables $X_i = (X_{i,1}, \ldots, X_{i,K})$ for each LSOA $i = 1, \ldots, N$. These data are not year or age-specific but describe socio-economic characteristics of the entire population of an LSOA measured at a specific point in time. Details about the predictive variables used in this study are provided in Section 2.2.

The mortality data, E_{ita} and D_{ita}, are available for calendar years 2001–2018 by single year of age. As the total exposure in any individual LSOA is very small, we will group ages for the construction of the mortality index. In this study, we will focus on three age groups: 60–69, 70–79 and 80–89.

Note that the boundaries of some LSOAs have changed during our observation period 2001–2018. All data used in this study are based on LSOA boundaries used in the 2011 census.

2.1. *Mortality data and relative risk*

In this study, we model the relative mortality risk in an individual LSOA $i \in \{1, \ldots, N\}$ compared to the average mortality in England. To define our measure of relative risk, we first define a baseline death rate m^b_{ta} for year t and age a for the whole of England in the usual way:

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

$$m^b_{ta} = \frac{\sum_{i=1}^N D_{ita}}{\sum_{i=1}^N E_{ita}}. \tag{2.1}$$

The model will be fitted using data from years $\mathcal{T}$ and age range $\mathcal{A}$. Without any additional information, the expected[1] total number of deaths $\hat{D}^0_i$ across all ages $a \in \mathcal{A}$ and years $t \in \mathcal{T}$ in LSOA i is given by

$$\hat{D}^0_i = \sum_{t\in\mathcal{T},\, a\in\mathcal{A}} m^b_{ta} E_{ita} \text{ for all } i = 1, \ldots, N,$$

and we define the observed relative risk of death R^0_i for an individual living in LSOA i as the ratio of the actual number of deaths to the expected number of deaths in that LSOA, that is,

$$R^0_i = \frac{\sum_{t\in\mathcal{T},\, a\in\mathcal{A}} D_{ita}}{\hat{D}^0_i} \text{ for all } i = 1, \ldots, N. \tag{2.2}$$

With our definition, the realised relative risk R^0 in any neighbourhood is a random variable since the realised number of deaths is random. In the following, we are interested in modelling the conditional expectation of R^0 given a vector of socio-economic characteristics.

Note that the relative risk R^0_i is not age and year specific. However, as mentioned above, we will calculate and model the relative risk using mortality data for different age ranges $\mathcal{A}$, see Section 5 for details.

2.2. Socio-economic characteristics

In Section 3, we will construct an index that explains differences in the mortality rates in different LSOAs based on differences in their socio-economic characteristics. In Wen (2022), a large universe of predictive variables for LSOA-specific mortality rates were considered. Based on findings there, we restrict our attention in this paper to 12 variables. They are listed in Table 1. Further details about those 12 variables, including data sources, can be found in the supplementary material published online.

The possible values of the first nine numerical variables $x_1, \ldots, x_9$ in Table 1 are on very different scales. For the purpose of visualisation, we standardise them to have mean zero and variance one. Details of the standardisation procedure can be found in the supplementary material.

Variable x_{10} is a categorical variable representing the urban–rural class of an LSOA and taking one of five values listed in Table 2.

In summary, the socio-economic characteristics of any neighbourhood are given as a vector taking values in the $K = 12$ dimensional space

$$L_0 = \mathbb{R}^9 \times \{1, \ldots, 5\} \times [0, 1]^2. \tag{2.3}$$

Note, that our urban–rural class indicator x_{10} distinguishes between urban conurbation in London and outside London. We have introduced that distinction as we found in previous research that mortality rates in London are rather different from mortality rates in other parts of England, see Cairns *et al.* (2021) and Wen (2022).

It is to be expected that the covariates in Table 1 are correlated. We report the empirical correlations in Table 3. We observe in Table 3 that there are some strong correlations, but we argue that none of the observed correlations is so strong that a variable should be removed. The few strong correlations we observe might be seen as problematic for parametric models as the parameters for individual covariates might not be identifiable when one variable can act as a proxy for another variable. However, since we have 32,844 observations, the inclusion of highly correlated variables is still meaningful and

[1] For the calculation of the expected number of deaths, we assume that for each age and calendar year (and sex) the number of deaths in any LSOA has a distribution such that the expectation is given by $m_{ita}E_{ita}$. The two most prominent examples for such distributions used for the number of deaths are the Poisson distribution and the negative binomial distribution. Without further information about LSOAs, we set $m_{ita} = m^b_{ta}$.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

Table 1. *Predictive variables used in our study to model the relative mortality risk, R^0. Variables $x_1, \ldots, x_9$ are standardised using a N(0,1) distribution function, x_{11} and x_{12} take values in [0, 1] and x_{10} is a categorical variable taking one of five values explained in Table 2.*

Predictive variable	Description
x_1	Old-age income deprivation
x_2	Employment deprivation (i.e. unemployment)
x_3	Proportion of the age-65+ population with no qualifications
x_4	Crime rate
x_5	Average number of bedrooms
x_6	Proportion of the population born in the UK
x_7	Wider barriers to housing (affordability, homelessness)
x_8	Employment/occupation: proportion in a management position
x_9	Proportion working more than 49 h per week (ages 16–74)
x_{10}	Urban–rural classification
x_{11}	Proportion of population aged 60+ in a care home with nursing care
x_{12}	Proportion of population aged 60+ in a care home without nursing care

Table 2. *Five categories for the urban–rural class (predictive variable X_{10}).*

Urban/Rural (UR) class	Definition
1	Urban conurbation (except London)
2	Urban city and town
3	Rural town and village
4	Rural hamlet and isolated dwellings
5	Urban conurbation (in London)

Table 3. *Correlations between the covariates, see Table 1 for details about the covariates. Empirical correlations have been calculated using all LSOAs in England and Wales regardless of their urban–rural classification. Note that x_{10} (urban–rural classification) is not included in the table.*

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{11}	x_{12}
x_1	1	0.79	0.73	0.62	–0.7	–0.26	0.7	–0.69	–0.53	0.02	0.05
x_2	0.79	1	0.77	0.57	–0.58	0.05	0.47	–0.8	–0.61	0	0.05
x_3	0.73	0.77	1	0.47	–0.5	0.1	0.4	–0.83	–0.66	0.02	0.04
x_4	0.62	0.57	0.47	1	–0.53	–0.34	0.61	–0.46	–0.38	–0.01	0.04
x_5	–0.7	–0.58	–0.5	–0.53	1	0.21	–0.63	0.43	0.37	0	–0.03
x_6	–0.26	0.05	0.1	–0.34	0.21	1	–0.59	–0.1	–0.01	–0.02	0.01
x_7	0.7	0.47	0.4	0.61	–0.63	–0.59	1	–0.41	–0.28	0	0.01
x_8	–0.69	–0.8	–0.83	–0.46	0.43	–0.1	–0.41	1	0.7	0.03	0
x_9	–0.53	–0.61	–0.66	–0.38	0.37	–0.01	–0.28	0.7	1	0.02	0.02
x_{11}	0.02	0	0.02	–0.01	0	–0.02	0	0.03	0.02	1	0.06
x_{12}	0.05	0.05	0.04	0.04	–0.03	0.01	0.01	0	0.02	0.06	1

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

significantly improves the fit. Additionally, in the random forest (RF) algorithm each decision only depends on one variable, and therefore this algorithm still works effectively when correlations are high – that is one of the reasons for choosing the RF estimator.

We would also argue that there is no surprise in the correlation table: for example x_1 is positively correlated with x_2, but negatively correlated with x_5 and x_8 meaning that the higher the level of income deprivation, the higher the level of employment deprivation and the smaller are the houses, and the fewer are working in management positions. For completeness, we also report correlation tables for individual urban–rural classes in the supplementary material published online. The general conclusions from those tables are similar to those obtained from correlations across all LSOAs. However, correlations with the more-minor variables x_3 to x_9 do vary more between different urban–rural classes.

3. The Longevity Index for England (LIFE)

As mentioned above, our aim in this paper is to construct an index that will explain the expected relative mortality risk in any neighbourhood based on the socio-economic characteristics of that neighbourhood. We call this index the Longevity Index for England, hereafter, the “LIFE index”.

3.1. Modelling relative mortality risk

As a starting point, we model the conditional expectation of the relative mortality risk R^0 given characteristics x:

$$f(x) := \mathbb{E}[R^0|x] \text{ for any } x \in L_0 \tag{3.1}$$

where $x = (x_1, \ldots, x_K)$ is the vector of predictive socio-economic variables taking values in a K-dimensional space L_0 of possible realisations of X_i. For the 12 variables in our empirical study, (see Table 1) L_0 is the 12-dimensional space defined in (2.3).

To estimate the regression function f, we will use a supervised machine learning algorithm called a RF, and we will denote this estimator of f by $\hat{f}^{\mathrm{RF}}$. Details of the estimation procedure are given in Section 4. Let us mention that other estimators for f could be used. For example, Wen (2022) compares the RF estimator with a local linear regression estimator.

3.2. Care homes

When we construct the LIFE index as an estimator for the relative risk in any LSOA, we need to take into account that in our sample of LSOAs there are some with care homes and some without care homes. Clearly, if a significant proportion of individuals in any LSOA are living in a care home, then this will increase the mortality rate in that LSOA and, therefore, increase the relative risk. However, this does not then properly reflect the main socio-economic characteristics of the LSOA $(x_1, \ldots, x_{10})$.

To offset this effect when constructing the LIFE index in the next section, we will make assumptions about the proportion of people living in a care home for any LSOA rather than using the actual proportion of people living in care homes in that LSOA. In other words, we are trying to answer the question: What would be the relative risk of dying in LSOA i if we kept all socio-economic variables to the values observed in that LSOA, but changed the proportion of people living in care homes to the average for the whole of England, or to some other chosen value.

3.3. The LIFE index

Based on the discussion so far, we now define our Longevity Index for England as the value of f for specific neighbourhoods using the socio-economic characteristics of this neighbourhood but replacing

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

the proportion of people living in care homes with the average for the whole of England. More precisely, we define the LIFE index for LSOA i as

$$R_i = f(\tilde{X}_i) \text{ with } \tilde{X}_i = \left(X_{i,1}, \ldots, X_{i,9}, X_{i,10}, \bar{X}_{11}, \bar{X}_{12}\right) \tag{3.2}$$

where $\bar{X}_{11}$ and $\bar{X}_{12}$ denote the average values of the proportion of an LSOA's population living in care homes with nursing and care homes without nursing, respectively. In replacing the true with the mean values of $\bar{X}_{11}$ and $\bar{X}_{12}$, it is helpful to note from Table 3 that x_{11} and x_{12} have a very low correlation with other socio-economic variables. First, this implies that care homes are not concentrated in neighbourhoods with particular socio-economic characteristics. Second, the lack of correlation means that when we replace $X_{i,11}$ and $X_{i,12}$ with their mean values, we do not need to alter the values of other predictive variables to compensate: that is, the presence of a care home does not artificially inflate or deflate an LSOA's other socio-economic, predictive variables.

Note that the index could be constructed with other adjustments to the care home variables. For example, we could choose to calculate the index based on setting $X_{i,11} = X_{i,12} = 0$. That would also be a good choice to model the relative mortality risk of the popoulation not living in care homes. However, we prefer the setting in (3.2) as we will use the index to calculate life expectancies as a function of socio-economic characteristics. Setting $X_{i,11} = X_{i,12} = 0$ would implicitly assume that no individuals will ever be in a care home, which is, of course, not reasonable. Additionally, base mortality m^b_{ta} incorporates excess care home deaths, and so we prefer to reflect this also in the index values.

Let us mention that our index is based on the conditional expectation in (3.1) and therefore, we can calculate the relative risk of dying for any values of the socio-economic variables, they do not need to be those observed in a specific LSOA. In other words, we can calculate the LIFE score for fictional neighbourhoods with specified socio-economic characteristics. This allows us to investigate how sensitive the relative mortality risk is with respect to changes in certain socio-economic variables. We will return to that point in Section 8.

From the construction of the LIFE index in (3.2), it is clear that the index is not explicitly age specific. Instead, it is an index summarising the socio-economic characteristics of all members of a small community regardless of their age. However, the LIFE index relies on an estimate of the relative risk function f in (3.2) that links socio-economic characteristics to death rates. The age range for death counts and exposures used to estimate the function f will of course have an impact on the obtained LIFE index values. We will investigate this further in Section 5.

4. Estimating the relative risk using the RF algorithm

As mentioned in Section 3, we can use a wide variety of nonparametric estimators for the regression function f in (3.1). In this study, we will use the RF algorithm. As this estimation step is at the heart of our index construction, we will explain our approach in detail.

4.1. Overview

Our RF algorithm consists of three stages. For each stage, we use R (R Core Team, 2021) and the R package *randomForest* (Liaw and Wiener, 2002).

4.1.1. Stage 1

The purpose of the first stage is exclusively to choose certain hyperparameters – we provide more details about the hyperparameters below. For this stage, we split (randomly) our data set, $\mathcal{S} = \{1, \ldots, N\}$, into two disjoint subsets:

- the training set, $\mathcal{S}^{\text{train}} \subset \mathcal{S}$ contains LSOAs used to "train" our model, that is, to choose optimal parameters determining $\hat{f}^{\text{RF}}$, see below for details; and

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

- the validation set, $\mathcal{S}^{\text{val}} \subset \mathcal{S}$ is the set of LSOAs used for model validation. In the first stage, data in the validation set are used for selecting hyperparameters; again, we explain details below.

Note that $\mathcal{S}^{\text{train}} \cap \mathcal{S}^{\text{val}} = \emptyset$ and $\mathcal{S}^{\text{train}} \cup \mathcal{S}^{\text{val}} = \mathcal{S}$. The parameter optimisation for the observations in the training set is repeated for each possible choice of hyperparameters. We then select hyperparameters for which $\hat{f}^{\text{RF}}$ produces the best fit for the observations in the validation set, and those values are then fixed for the hyperparameters in the second stage.

4.1.1. Stage 2

In the second stage, we split our complete data set again into two subsets allocating LSOAs randomly (and independent of the allocation in the first stage) to either:

- the training set, $\mathcal{S}^{\text{train}} \subset \mathcal{S}$ containing LSOAs used to choose optimal parameters determining $\hat{f}^{\text{RF}}$ using the optimal hyperparameters determined in stage one; or
- the test set, $\mathcal{S}^{\text{test}} \subset \mathcal{S}$ which is a subset of LSOAs that are only used for evaluating how good our estimated function $\hat{f}^{\text{RF}}$ (fitted to data in $\mathcal{S}^{\text{train}}$) can predict the relative risk in out of sample LSOAs.

Using the values for the hyperparameters obtained in the first stage, we fit our estimator $\hat{f}^{\text{RF}}$ to the observations in our new training set and evaluate the goodness-of-fit using the test set.

So, both stages follow the same idea, but in the first we refit $\hat{f}^{\text{RF}}$ to the stage-one training set many times to choose optimal hyperparamters, while in stage two $\hat{f}^{\text{RF}}$ is fitted to the stage-two training set only once to assess the out-of-sample goodness-of-fit.

In this paper, we split the set of all $N = 32,844$ LSOAs into two equally sized disjoint sets to obtain a training set and a validation or test set in the two stages.

4.1.3. Stage 3

In the final stage of the estimation procedure, we run the RF algorithm with all chosen hyperparameters for the full set of $N = 32,844$ LSOAs to produce the final estimate of the regression function f in (3.1) and obtain the LIFE index values from (3.2).

4.2. Fitting a single tree

A RF consists of $B > 1$ regression trees also known as decision trees. We will here briefly discuss how each tree is constructed as a crude estimator of the regression function f in (3.1). In the next section, we will then turn to combining many trees into a RF.

For fitting an individual tree with index $b \in \{1, \ldots, B\}$, we only use a subset $\mathcal{S}^b$ of the LSOAs in the training set $\mathcal{S}^{\text{train}}$ in both stages. The procedure for growing a tree is the same for stages one and two. The choice of $\mathcal{S}^b$ is explained in Section 4.3.

Constructing an individual tree is an iterative procedure. We start with defining our initial estimator $\hat{f}_0^{(b)}$ as the average of all observed values of the relative risk R^0 of the LSOAs in the set $\mathcal{S}^b$, that is,

$$\hat{f}_0^{(b)}(x) = \frac{1}{|\mathcal{S}^b|} \sum_{i \in \mathcal{S}^b} R_i^0 \qquad \text{for all } x \in L_0$$

where $|\mathcal{S}^b|$ is the size of the data set $\mathcal{S}^b$.

In the next step, we choose one explanatory variable, say x_{k^*}, and a level l^*, and split the initial node L_0 into the two disjoint subsets:

$$L_{1,1}^b = \{x \in L_0 : x_{k^*} < l^*\} \tag{4.1}$$

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

$$L_{1,2}^{b} = \{x \in L_0 : x_{k^*} \geq l^*\} \tag{4.2}$$

This procedure is now repeated but, in addition to choosing an explanatory variable x_{k^*} and a threshold level l^*, we also choose one of the sets (nodes) $L_{1,1}^{b}$ and $L_{1,2}^{b}$ which we then split in the next step.

Starting with $s = 1$ and the two nodes defined in (4.1) and (4.2), we now apply the following iterative procedure:

- Choose one subset L_{s,j^*}^{b} ($j^* \in 1, \ldots, s+1$) out of the $s+1$ subsets formed by the first s splits. Also, choose an explanatory variable x_{k^*} and a threshold l^*
- Split L_{s,j^*}^{b} into two subsets and leave all other subsets unchanged.

With this procedure, we have the following nodes available after $s+1$ splits:

$$L_{s+1,j}^{b} = L_{s,j}^{b} \text{ for } j = 1, \ldots, s+1,\ j \neq j^* \tag{4.3}$$

$$L_{s+1,j^*}^{b} = \{x \in L_{s,j^*}^{b} : x_{k^*} < l^*\} \tag{4.4}$$

$$L_{s+1,s+2}^{b} = \{x \in L_{s,j^*}^{b} : x_{k^*} \geq l^*\} \tag{4.5}$$

Equation (4.3) states that split $s+1$ does not affect any nodes other than L_{s,j^*}^{b}. Equations (4.4) and (4.5) mean that all LSOAs with characteristics x in node L_{s,j^*}^{b} for which $x_{k^*} \geq l^*$ are put into a new node $L_{s+1,s+2}^{b}$ so that only those LSOAs with characteristics in L_{s,j^*}^{b} and $x_{k^*} < l^*$ remain in that node and are then contained in L_{s+1,j^*}^{b} after $s+1$ splits.

We now define the estimator $\hat{f}_s^{(b)}(x)$ of the regression function f in (3.1) obtained from one tree b after splitting L_0 into $s+1$ nodes as

$$\hat{f}_s^{(b)}(x) = \sum_{j=1}^{s+1} r_j \mathbb{I}_{(x \in L_{s,j}^{b})}$$

where r_j is the average observed relative risk R^0 for LSOAs X_i in node j, that is,

$$r_j = \frac{1}{\sum_{i=1}^{N} \mathbb{I}_{(X_i \in L_{s,j}^{b})}} \sum_{i=1}^{N} \mathbb{I}_{(X_i \in L_{s,j}^{b})} R_i^0 \qquad \text{for all } j = 1, \ldots, s+1.$$

As explained above, for each new split, we need to choose an existing node L_{s,j^*}^{b}, an explanatory variable x_{k^*} and a threshold l^*. Those are chosen such that the fit of the new estimator $\hat{f}_{s+1}^{(b)}(x) = \hat{f}_{s+1}^{(b)}(x; j^*, x_{k^*}, l^*)$ to the observed values of the relative risk, R_i^0 for $i \in \mathcal{S}^b$, is optimised. More specifically, our choice minimises the residual sum of square

$$\text{RSS}_s^{b}(j, x_n, l) = \sum_{i \in \mathcal{S}^b} \left(R_i^0 - \hat{f}_{s+1}^{(b)}(X_i; j, x_n, l) \right)^2 \tag{4.6}$$

$$\left(j^*, n^*, l^*\right)_s^b = \underset{j,n,l}{\text{argmin}}\, \text{RSS}_s^{b}(j, x_n, l) \tag{4.7}$$

Within the RF algorithm, for each split, rather than optimise over all $K = 12$ of the predictive variables, we optimise over a subset of m variables. This subset is chosen randomly for each new split (see, for example, James *et al.*, 2013). The purpose of this is to increase the variability and reduce the correlation between individual trees. Without restricting the set of variables we find that trees are very similar. The parameter m is a hyperparameter, and we explain its choice in Section 4.4.

Finally, we stop splitting nodes further as soon as any obtained node contains less than M observations where M is another hyperparameter. In our empirical study, we choose $M = 200$ as that choice achieves a good balance between goodness-of-fit and overfitting. Díaz-Uriarte and Alvarez de Andrés

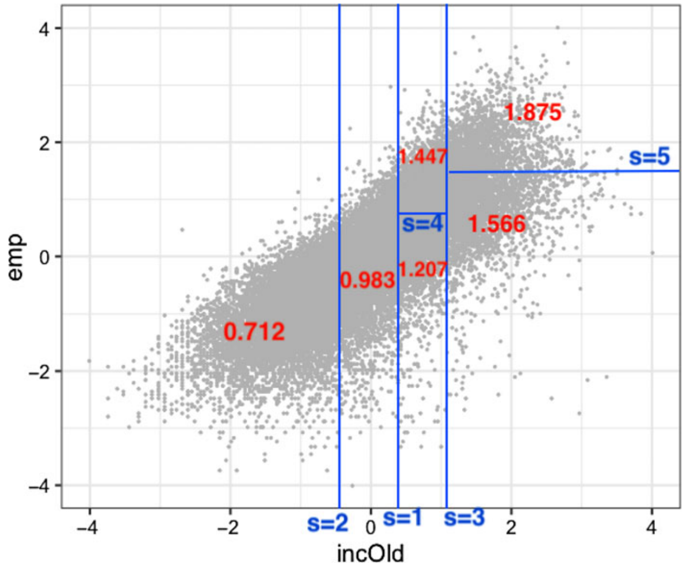

Figure 1. The values of the piecewise constant regression tree $\hat{f}_5^{(b)}(x)$ in (4.8) after five splits. Blue solid lines show the boundaries of the six nodes. Red numbers are the estimated relative risk for all LSOAs in each of the those nodes. Each of the gray dots represents the observed values of old-age income deprivation and employment deprivation for a single LSOA in the training set for this example.

(2006) provide a general discussion of the minimum node size M, principles of choosing it, and examples on its potential impact on the model performance and computation time. Wen (2022) has a more detailed discussion around M in the specific context of modelling relative mortality risk at neighbourhood level using the RF algorithm, including its impact on the complexity of underlying trees and the standard deviation of the outcomes produced by individual trees. Although the model's out of sample performance does not appear to be very sensitive to the choice of M, choosing M to be 200 rather than, say, 5 or 50 significantly saves computation time without sacrificing the predictive power of the RF estimator in our application.

In Figure 1, we illustrate the construction of one tree using our data set of LSOAs. In this example, only two variables are considered for potentially splitting nodes: old-age income deprivation, x_1 and employment deprivation, x_2. In total, $s = 5$ splits have been performed leaving us with $s + 1 = 6$ nodes, and our estimator for f is given by

$$\hat{f}_5^{(b)}(x) = \begin{cases} 0.7117 \ \text{ for } \ x \in L_{5,1}^b = \{x : x_1 < -0.418\} \\ 0.9830 \ \text{ for } \ x \in L_{5,2}^b = \{x : -0.418 \le x_1 < 0.351\} \\ 1.2070 \ \text{ for } \ x \in L_{5,3}^b = \{x : 0.351 \le x_1 < 1.06 \text{ and } x_2 < 0.701\} \\ 1.4470 \ \text{ for } \ x \in L_{5,4}^b = \{x : 0.351 \le x_1 < 1.06 \text{ and } 0.701 \le x_2\} \\ 1.5660 \ \text{ for } \ x \in L_{5,5}^b = \{x : 1.06 \le x_1 \text{ and } x_2 < 1.441\} \\ 1.8750 \ \text{ for } \ x \in L_{5,6}^b = \{x : 1.06 \le x_1 \text{ and } 1.441 \le x_2\} \end{cases} \tag{4.8}$$

In Figure 2, we show the order of the performed splits: the first three splits are all based on x_1 (old-age income deprivation). Only after three splits using x_1, two of the obtained nodes are split using x_2 (employment deprivation). This clearly shows, that for this example x_1 has a higher explanatory power than x_2 since splitting the early nodes in our tree according to the old-age income deprivation score reduces the residual sum of squares RSS more than early splits with respect to employment deprivation would achieve.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

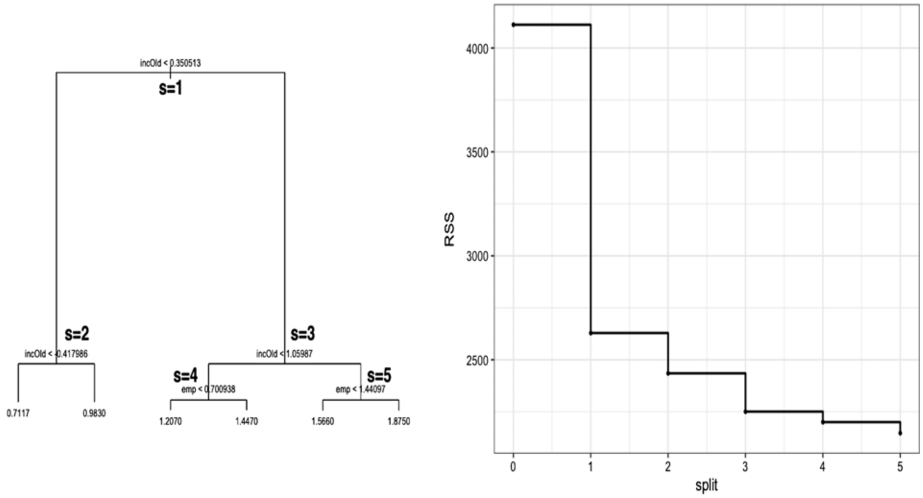

Figure 2. *A graphical representation of our example regression tree $\hat{f}_5^{(b)}(x)$ in (4.8), and the residual sum of squares, RSS$_s^b$ (4.6) as a function of the number of splits s for this tree.*

To illustrate this specific example further, we also report the residual sum of squares, RSS_s^b, in Figure 2. As expected, we find that the early splits result in the greatest reduction of RSS_s^b.

4.3. Many trees form a RF

Having seen how an individual regression tree is fitted to observations in a set $\mathcal{S}^b$, we now turn to describing how we choose the sets $\mathcal{S}^b$ and how we combine many trees to obtain our final estimator $\hat{f}^{\mathrm{RF}}$ for the regression function f in (3.1).

For each tree $b \in \{1, \ldots, B\}$, the set $\mathcal{S}^b$ is obtained by (see, for example, James *et al*., 2013)

1. sampling randomly with replacement from the training data set $\mathcal{S}^{\mathrm{train}}$ to obtain a sample of the same size as $\mathcal{S}^{\mathrm{train}}$, and then
2. removing all duplicates from that sample.[2]

Repeating this procedure B times, we obtain B subsets $\mathcal{S}^b \subseteq \mathcal{S}^{\mathrm{train}}$ of the training data set.

To introduce more randomness in the construction of the RF estimator $\hat{f}^{\mathrm{RF}}$ we also, as remarked before, restrict the predictive variables considered at each split in any individual tree. Rather than choosing a predictive variable x_k out of all p variables when minimising the residual sum of squares RSS_s^b in (4.6), we follow James *et al*. (2013) and choose x_k from a subset of m predictive variables. As mentioned in Section 4.2, this subset of predictive covariates is randomly chosen for each split within each tree.

So, each tree $b = 1, \ldots B$ is fitted to a randomly chosen subset $\mathcal{S}^b$ of observations from the training data set, and RSS_s^b is optimised with respect to m randomly chosen predictive variables. In this way, we obtain a total of B regression functions $\hat{f}^{(b)}$. The number m of predictive variables considered for each split of nodes is a hyperparameter and will be chosen in stage one using cross-validation. As mentioned earlier, m is then fixed in stage two.

Our final RF estimator $\hat{f}^{\mathrm{RF}}$ for the regression function f in (3.1) is obtained by taking the average over all individual regression trees $\hat{f}^{(b)}$, that is,

$$\hat{f}^{\mathrm{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{(b)}(x) \text{ for any } x \in L_0 \tag{4.9}$$

[2] If we generate a random sample of size n with replacement from a set also of size n, and then remove duplicates, we are left with a sample with a random size and mean size equal to $n(1 - n^{-1})^2 \approx n(1 - e^{-1}) \approx 0.632n$.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

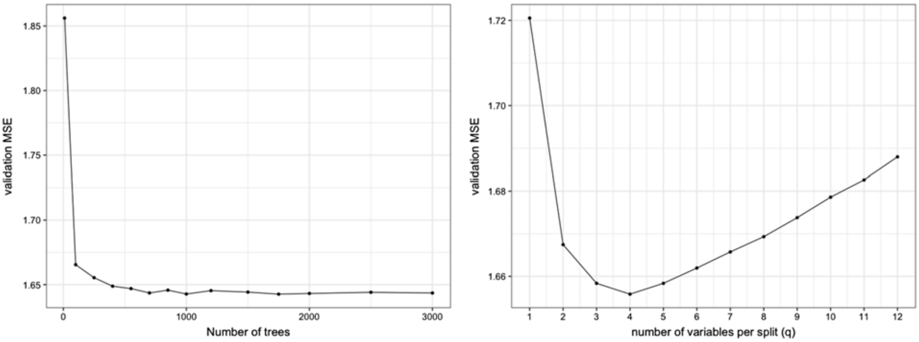

Figure 3. *Validation MSE calculated following (4.10) and over LSOAs in the validation set $\mathcal{S}^{va}$, of a random forest model trained using the relative risk of England males aged 70–79, and with different numbers of trees B (left plot, with m set as 4) and different numbers of variables considered per split m (right plot, with B set as 2500). The 12 predictive variables outlined in Table 1 are used.*

Note that $\hat{f}^{\text{RF}}$ is piecewise constant over the full range of values of $x \in L_0$ as it is an average over a finite number of piecewise constant regression tree functions $\hat{f}^{(b)}$. However, $\hat{f}^{\text{RF}}$ can take many more values compared to any individual tree $\hat{f}^{(b)}$.

4.4. Hyperparameter selection (stage 1)

With the minimum node size, $M = 200$, fixed, the regression function $\hat{f}^{\text{RF}}$ in (4.9) will depend on two further hyperparameters: the number m of predictive variables considered for each split, and the number of trees, B. Both of those parameters need to be chosen. One could argue that we can also choose the size N^{train} of the training set $\mathcal{S}^{\text{train}}$, but, for simplicity, we choose that set to include half of the available observations with the other half being included in the validation set $\mathcal{S}^{\text{val}}$ in stage one. For the $N = 32,844$ LSOAs in our empirical study, we clearly have that both sets include $N^{\text{train}} = N^{\text{val}} = 16,422$ LSOAs.

The hyperparameters B and m are chosen in stage one in the following way: we fit $\hat{f}^{\text{RF}} = \hat{f}^{\text{RF}}_{B,m}$ to the data in the training set using different values for B and m. For each combination (B,m) considered, we then evaluate the fit of the obtained estimate $\hat{f}^{\text{RF}}_{B,m}$ to the data in the validation set using the mean squared error as criterion.

$$\text{MSE}(B,m) = \frac{1}{N^{\text{val}}} \sum_{i \in \mathcal{S}^{\text{val}}} \frac{\left(R_i^0 - \hat{f}^{\text{RF}}_{B,m}(X_i)\right)^2}{\hat{f}^{\text{RF}}_{B,m}(X_i)/\hat{D}_i^0} = \frac{1}{N^{\text{val}}} \sum_{i \in \mathcal{S}^{\text{val}}} \frac{\left(D_i - \hat{D}_i^{RF}\right)^2}{\hat{D}_i^{RF}} \tag{4.10}$$

where $D_i = R_i^0 \hat{D}_i^0 = \sum_{t,a} D_{ita}$ is the observed total number of deaths across all ages a and years t in LSOA i, and $\hat{D}_i^{RF} = \hat{f}^{\text{RF}}_{B,m}(X_i)\hat{D}_i^0$ is the expected total number of deaths adjusted with the fitted relative risk $\hat{f}^{\text{RF}}_{B,m}(X_i)$ for LSOA i.

In Figure 3, we plot MSE(B, m) in (4.10) for different values of B (with $m = 4$) and different values of m (with $B = 2500$). The figure shows that the out-of-sample performance of the RF is not worsening as more trees are grown, and we choose $B = 2500$ as we think this will be a good compromise between computational effort and goodness-of-fit. However, we find that considering $m = 4$ predicative variables in (4.6) leads to the smallest mean squared error.

Table 4 summarises our choice of hyper-parameters.

4.5. Goodness-of-fit (stage 2)

In order to assess the out-of-sample performance of the proposed RF estimator applied to the mortality data for the 32,844 LSOAs, we move on to stage two as mentioned in Section 4.1.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

Table 4. *Settings of the final random forest model we use for creating the mortality index for England males.*

Parameter/hyperparameter	Notation	Value
Number of trees	B	2500
Total number of variables	p	12
Number of variables to consider per split	m	4
Minimum size of terminal nodes (floor)	M	200

Table 5. *Test set MSE of the proposed random forest model fitted to three randomly chosen training sets (rounds) for data from different age groups. The applied hyperparameters are listed in Table 4.*

Age	Round 1	Round 2	Round 3	Overall
60–69	1.273	1.253	1.299	1.275
70–79	1.686	1.689	1.688	1.688
80–89	2.799	2.744	2.935	2.826

To this end, we randomly split the set of all 32,844 LSOAs into two equally sized subsets: the training set $\mathcal{S}^{\text{train}}$ (different from before) and the test set $\mathcal{S}^{\text{test}}$. The training set $\mathcal{S}^{\text{train}}$ chosen at this stage is a random sample of $\mathcal{S}$ and independent of the training set chosen in stage 1. The sets $\mathcal{S}^{\text{train}}$ and $\mathcal{S}^{\text{test}}$ are disjoint and $\mathcal{S}^{\text{train}} \cup \mathcal{S}^{\text{test}} = \mathcal{S}$. We then construct an estimator $\hat{f}^{\text{RF}}$ using the data in $\mathcal{S}^{\text{train}}$ and the hyper-parameters in Table 4. To quantify the goodness-of-fit of the obtained estimator $\hat{f}^{\text{RF}}$, we evaluate its out-of-sample fit to data in the test set by calculating the mean squared error as in (4.10) but now considering LSOAs in the test set rather than the validation set, that is,

$$\text{MSE}^{\text{test}} = \frac{1}{N^{\text{test}}} \sum_{i \in \mathcal{S}^{\text{test}}} \frac{\left(R_i^0 - \hat{f}^{\text{RF}}(X_i)\right)^2}{\hat{f}^{\text{RF}}(X_i)/\hat{D}_i^0} \tag{4.11}$$

Clearly, the realised values of MSE^{test} will depend on the randomly chosen LSOAs in $\mathcal{S}^{\text{train}}$ and $\mathcal{S}^{\text{test}}$. To get an idea of how sensitive the results are to the randomised choice of $\mathcal{S}^{\text{train}}$ and $\mathcal{S}^{\text{test}}$, we calculate MSE^{test} for three different splits (rounds) of our data into training and test sets. The results are provided in Table 5 for data based on different age ranges. We find that there is not much variation in the values of MSE^{test} between rounds. We also see that the goodness-of-fit of $\hat{f}^{\text{RF}}$ is much better when it is estimated from mortality data at younger ages.

4.6. Robustness

The rather small variation of the test set MSEs over different randomly chosen training sets in Table 5 is an indication that the fitted relative risk, and therefore the LIFE index, is a robust estimator of the true underlying relative mortality risk. To investigate robustness further we now split the annual data for the observation period 2001–2018 into two subsets: data for even years 2002, 2004, . . . and data for odd years 2001, 2003, . . . This split leaves us with two subsets each consisting of nine years of observations. We chose to split the observation period in this way to avoid any impact of potential trends in the relative mortality risk over time.

We now apply the above methods to obtain estimates $\hat{f}^{\text{RF}}(x)$ with data from only one of the two observation subsets and then compare the results.

We present scatter plots of the estimated values of $\hat{f}^{\text{RF}}(x)$ based on odd years (horizontal axis) and even years (vertical axis) using mortality data for different age ranges in Figure 4. The plots clearly

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

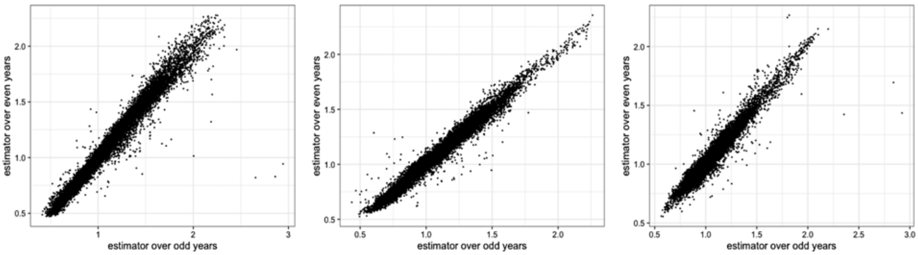

Figure 4. *Estimated relative risk over 16,422 LSOAs in $\mathcal{S}^{te}$ by the random forest model trained using LSOAs in $\mathcal{S}^{tr}$ and relative risks of the two year groups, $R^{0,odd}$ and $R^{0,even}$. x-axis: model trained with odd years; y-axis: model trained with even years. Left to right: relative risk of age 60–69, 70–79 and 80–89.*

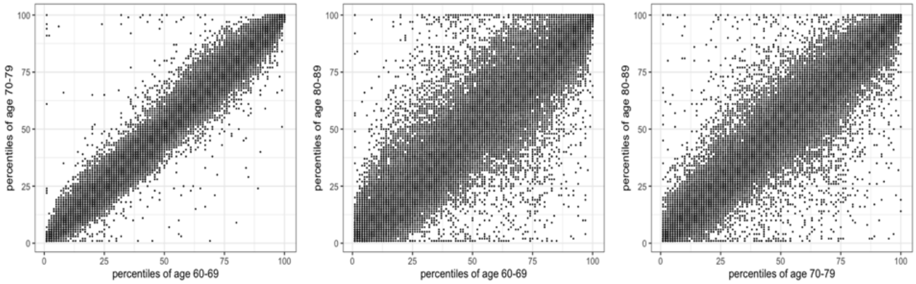

Figure 5. *Percentiles of LIFE index values R_i compared between indices estimated from different age groups over all 32,844 LSOAs. Left: 60–69 versus 70–79; Middle: 60–69 versus 80-89; Right: 70–79 versus 80–89.*

show that the estimated values of the relative risk for individual LSOAs are very similar when mortality data from different years are used, in particular, there seems to be no systematic differences – this is further evidence that the results of our RF estimator are robust. Any variation we see is most likely due to sampling variation in the deaths counts rather than systematic differences.

4.7. Final index values (stage 3)

The hyperparameters have been chosen in stage 1 and goodness-of-fit and robustness assessed in stage 2, and it has been concluded that the RF algorithm has produced a good estimate of f with the chosen parameters. In the final stage, we simply rerun the RF algorithm but, instead of using only half of the data, we use the full set of 32,844 LSOAs.

5. Fitting the LIFE index to different age groups

As mentioned in Section 3.3, the LIFE index is not directly age specific. However, its estimated values depend upon the specified age range, $\mathcal{A}$, and the index values obtained can be assumed to apply to either the whole of that age range or to the midpoint of that range.

To investigate the effect of different age groups on the estimated LIFE index value, we compare index values obtained from fitting $\hat{f}^{\mathrm{RF}}(x)$ to mortality data for three age groups: 60–69, 70–79 and 80–89. We report Q–Q-plots of the obtained index values R_i (Equation (3.2)) for all 32,844 LSOAs in Figure 5.

Figure 5 shows a very strong dependency between the LIFE index for age groups 60–69 and 70–79 (left-hand plot). This dependency is slightly weaker when we compare the age group 80–89 with the younger ages (a greater spread of points in the middle and right hand plots) but the dependency is still strong. Wen (2022) found similar results for an index constructed using local linear regression.

Table 6. *Spearman's rank correlation of IMD and LIFE index values. The LIFE index has been fitted to the mortality experience in different age groups.*

Age group	40–49	50–59	60–69	70–79	80–89
Correlation with IMD	0.941	0.932	0.917	0.896	0.836

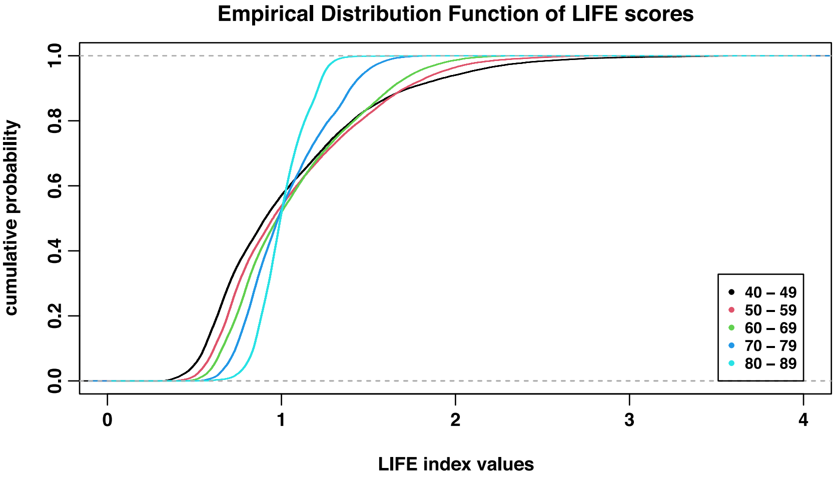

Figure 6. *Cumulative distribution function of the LIFE scores for all 32,844 LSOAs fitted to mortality data for different age groups.*

Empirical distributions of the R_i are plotted in Figure 6 for different age groups. We can see that variation in relative mortality risk is greater for younger ages than older ages, an observation that is consistent with previous research on socio-economic variation in mortality in various populations (see, for example, Mackenbach *et al*., 2003; Mackenbach *et al*., 2015; Chetty *et al*., 2016; Wen *et al*., 2020 and Wen *et al*., 2021 and references therein).

6. The LIFE index versus the IMD

The IMD is published by the Department for Communities and Local Government in the UK (Smith *et al*., 2015). The IMD is designed as a general measure for deprivation. The LIFE index on the other hand has been produced specifically as a measure of mortality deprivation with the aim to predict mortality differences between LSOAs. We would expect the two indices to have a highrank correlation. To check this hypothesis, we plot the LIFE index values fitted to mortality data for ages 40–49, 60–69 and 80–89 versus the IMD scores in Figure 7. We also report Spearman's rank correlations between the LIFE index fitted to different age ranges and the IMD scores in Table 6.

We find that the rank correlation is indeed high, in particular, when the LIFE index is fitted to mortality data at younger ages, see Table 6. We also observe in Figure 7 that there is a strong dependency between the scores of the two indices, but that there are some outlier LSOAs with a rather low IMD score (little deprivation) and a rather high relative mortality risk.

We did consider the outlier LSOAs, for example, for ages 60–69 in some detail. Individual predictive variables for these LSOAs tended to be towards the tails of the data but not too extreme, but, in higher dimensions, the vectors X_i for these LSOAs were clearly positioned around the fringes of the cluster of observations for the 32,844 LSOAs. As with most regression methods, estimates at the edges of a dataset do carry higher levels of uncertainty than estimates in the middle of the dataset.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

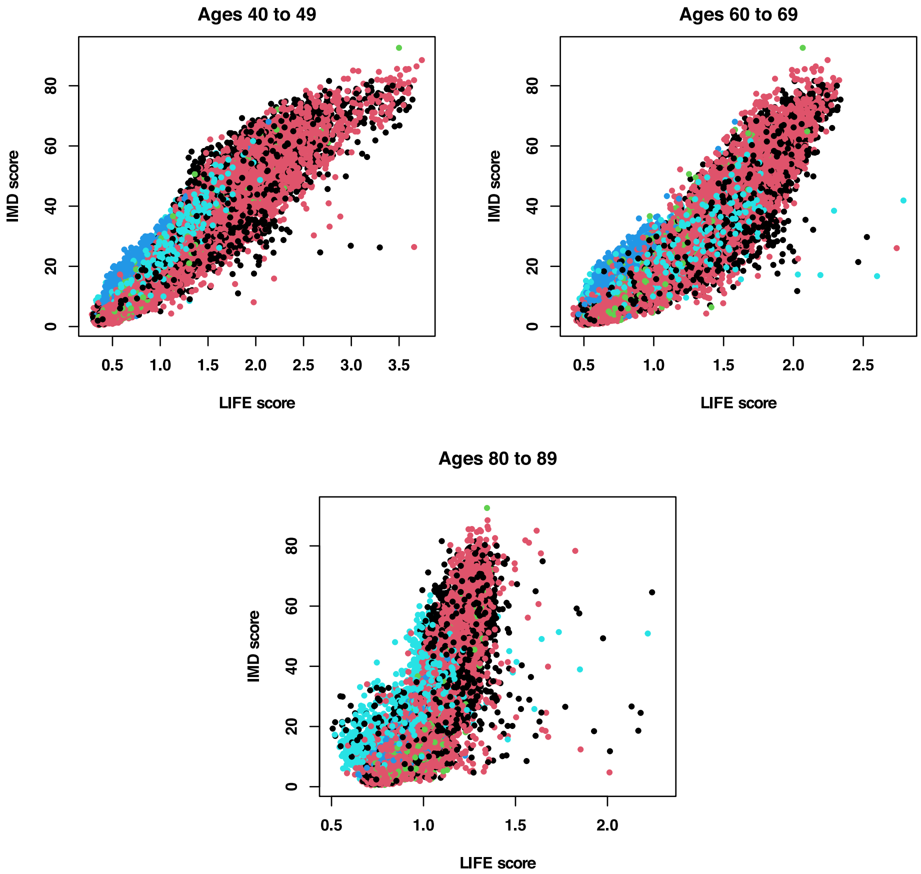

Figure 7. *Scatterplot of LIFE index versus IMD. The LIFE index is based on an estimated relative mortality risk fitted to mortality data for ages 40–49 (top left), 60–69 (top right) and 80–89 (bottom). Colour indicate the urban–rural class of an LSOA: conurbations (black), cities/towns (red), villages (green), rural areas (dark blue) and London (light blue).*

The colouring of the dots in Figure 7 reveals how significantly the inclusion of urban–rural class has impacted on estimates of the relative risk compared to the IMD, particularly ages 40–49 and 60–69. In the upper plots, the dark blue dots representing very rural areas are mostly shifted to the left of the main diagonal. This indicates that the inclusion of urban–rural class in the RF model estimates significantly lower mortality in these rural areas than would be suggested by the IMD, which takes no explicit account of urban–rural class. Indeed the IMD includes a subdomain called geographical barriers which counts greater distance to services as meaning an area is more deprived. But in mortality terms (at least at the macro scale) the opposite is true: larger, more-rural or otherwise less-dense LSOAs have lower mortality even though one has to travel further for essential services.

The RF model also predicts lower mortality for London (light blue dots) than the IMD predicts. In this case, the reason is less clear but needs further investigation: what is missing in the IMD that is to the advantage of London and to the disadvantage of other areas, particularly other conurbations and cities.

7. Distribution of low- and high-risk groups across urban–rural classes

In this section, we investigate if there are differences between urban–rural classes in the number of low- and high-risk mortality populations. More specifically, we denote by q^R_α the empirical α-quantile of the

Table 7. *Distribution of LSOAs across urban–rural classes for different subpopulations. The numbers in brackets refer to the proportion (in %) of all LSOAs in a group that fall within an urban–rural class. The subpopulation groups are defined in (7.1). See Table 2 for the definition of urban–rural classes.*

		Relative risk R_k fitted to ages 60–69			
UR class	All LSOAs	$G^l_{0.05}$	$G^u_{0.05}$	$G^l_{0.5}$	$G^u_{0.5}$
1	7921	220 (2.8)	944 (11.9)	2681 (33.9)	5240 (66.2)
2	14,515	734 (5.1)	610 (4.2)	7205 (49.6)	7310 (50.4)
3	3056	162 (5.3)	8 (0.3)	2216 (72.5)	840 (27.5)
4	2542	326 (12.8)	0 (0.0)	2444 (96.1)	98 (3.9)
5	4810	201 (4.2)	81 (1.7)	1876 (39.0)	2934 (61.0)
Total	32,844	1643 (5.0)	1643 (5.0)	16,422 (50.0)	16,422 (50.0)
		Relative risk R_k fitted to ages 80–89			
1	7921	223 (2.8)	597 (7.5)	2558 (32.3)	5363 (67.7)
2	14,515	657 (4.5)	768 (5.3)	6979 (48.1)	7536 (51.9)
3	3056	129 (4.2)	86 (2.8)	1863 (61.0)	1193 (39.0)
4	2542	111 (4.4)	32 (1.3)	1926 (75.8)	616 (24.2)
5	4810	523 (10.9)	160 (3.3)	3096 (64.4)	1714 (35.6)
Total	32,844	1643 (5.0)	1643 (5.0)	16,422 (50.0)	16,422 (50.0)

estimated relative risk R^0 in all LSOAs, and we then define four groups of LSOAs: the lower 5% and 50% quantile groups, and the upper 5% and 50% quantile groups,

$$G^l_\alpha := \left\{k : R^0_k < q^R_\alpha\right\} \text{ and } G^u_\alpha := \left\{k : R^0_k > q^R_{1-\alpha}\right\} \text{ for } \alpha = 0.05, 0.5 \tag{7.1}$$

where q^R_α is the empirical α quantile of the fitted values R^0_k.

Table 7 shows how urban–rural classes are distributed in each of those groups.

An interesting result in Table 7 is that out of the 2542 LSOAs that are classified as rural hamlets and isolated dwellings (urban–rural class 4) none can be found in the high-risk group $G^u_{0.05}$ when the relative risk is fitted to mortality data for ages 60–69. Similarly, only eight of the 3056 LSOAs in urban–rural class 3 are found in the high-risk group. On the other hand, 11.9% of large conurbations outside London made it into the top 5% risk group while only 1.7% of LSOAs in London are in that group. The data in Table 7 clearly show the strong impact that the urban–rural class has on the estimated relative mortality risk of an LSOA with the general conclusion that LSOAs in large cities tend to have a higher mortality risk than LSOAs in rural areas. London is an exception for the very high-risk group but we also find that more than half (61%) of LSOAs in London have a mortality risk greater than the median for England.

The picture changes slightly when we consider the oldest age group, 80–89, in our data set. However, the general conclusion seems to be unchanged: large cities have higher mortality than rural areas.

8. Impact of specific variables

Since the LIFE index is based on a non-parameteric estimator of the regression function f in (3.1), it is not straightforward to assess the impact of specific variables on the index value from the sign or magnitude of specific parameters as is often possible for parametric regression models. Instead, we study the values of $\hat{f}^{\mathrm{RF}}(x)$ for a certain range of covariate values where we vary only some variables while leaving others constant.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

In Wen (2022), it was found that employment and old-age income deprivation are two of the most significant predictors of mortality rates. We therefore focus on those two variables first. We also include here x_6 – the proportion of an LSOA's population born in the UK – as an example of a less important variable. Figure 8 shows the fitted relative mortality risk as a function of those three covariates. In each of the plots, all variables except the variable on the horizontal axis have been set to their median calculated across all LSOAs. More specifically, for the first row in Figure 8 (old-age income deprivation) the value shown for LSOA i is calculated as

$$f\left(X_{i,1}, X_2^{50}, \ldots, X_{11}^{50}, X_{i,12}^{50}\right) \tag{8.1}$$

where x^{50} denotes the empirical 50% quantile of covariate x. We use a similar approach for X_2 and X_6. This allows us to zoom in on the specific effect of one variable.

We can clearly see that old-age income deprivation and employment deprivation have a similar effect on the risk of dying with high levels of deprivation associated with high levels of mortality. However, comparing the range of risk values (y-axis), we find that income deprivation is a much better variable than employment deprivation to distinguish between low- and high-risk LSOAs. This is particularly true when data for ages 60–69 are used to fit the relative risk function. Not surprisingly, old-age income deprivation is still a good variable to explain differences at the older ages 80–89, but employment in an LSOA has little explanatory power for mortality differences in that age group.

Turning to x_6 – the proportion of the population born in the UK – this variable has very limited explanatory power (a narrow range of relative risk values) when all other variables are set to the median. Nevertheless, Wen (2022) has found that it is a variable that helps the RF algorithm to better predict observed mortality risk. While we find that higher numbers of UK-born residents seems to slightly increase mortality in an LSOA, this effect is relatively small. We also observe in Figure 8 that the relative risk is below one for all LSOAs. This is clearly a consequence of setting all other covariates to the median. The relative risk values and our conclusions about x_6 might change when other covariates are set to different values rather than the median. Our proposed non-parametric estimation of the relative risk would allow for such more detailed empirical studies but that is beyond the scope of this paper. Also of note is the fact that the steep portion of both plots for x_6 is well to the left of the median ($x_6 = 0$). A potential reason for this is how x_6 interacts with other predictive variables. It is only for the 20% lowest where there is a significant dependency between x_6 and other predictive variables. For the upper 80% of the distribution of x_6, there is very little dependency with other variables.

Finally, we conclude from Figure 8 that the impact of the three considered variables is comparable in all five urban–rural classes.

The LIFE index has been constructed by keeping all variables at their observed levels except the proportion of residents living in care homes, x_{11} and x_{12}, see Section 3.3 for details. To investigate further the impact of the three variables, we now plot the values of the LIFE index for all LSOAs as a function of one covariate in Figure 9.

We find in Figure 9 that, as expected, there is a lot more fluctuation when the variability from other covariates is not removed as in Figure 8. We can see that old-age income deprivation has the strongest correlation with estimated relative mortality risk followed by employment deprivation and, finally, there is no clear pattern linking the population born in the UK to the relative risk. In each plot, the colour of each point shows the urban–rural class. We can see that a London-effect is clearly visible: London has lower mortality as seen from the plot on the top left, and London has also a relatively large population of people born abroad which is clearly visible in the lower plot.

Finally, we consider the joint impact of two variables on the LIFE index. In Figure 10, we show heat plots of LIFE index values for ages 60–69 as functions of old-age income deprivation, x_1, and employment deprivation, x_2. All variables $x_3, \ldots, x_9$, x_{11} and x_{12} are set to their median and the three panels show urban–rural classes 1, 4 and 5 from left to right. It is notable that, in all three cases, for less-deprived LSOAs, the divisions between bands of colour are nearly vertical indicating that old-age income deprivation is the main driver of the LIFE index out of the two variables. However, as we move

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

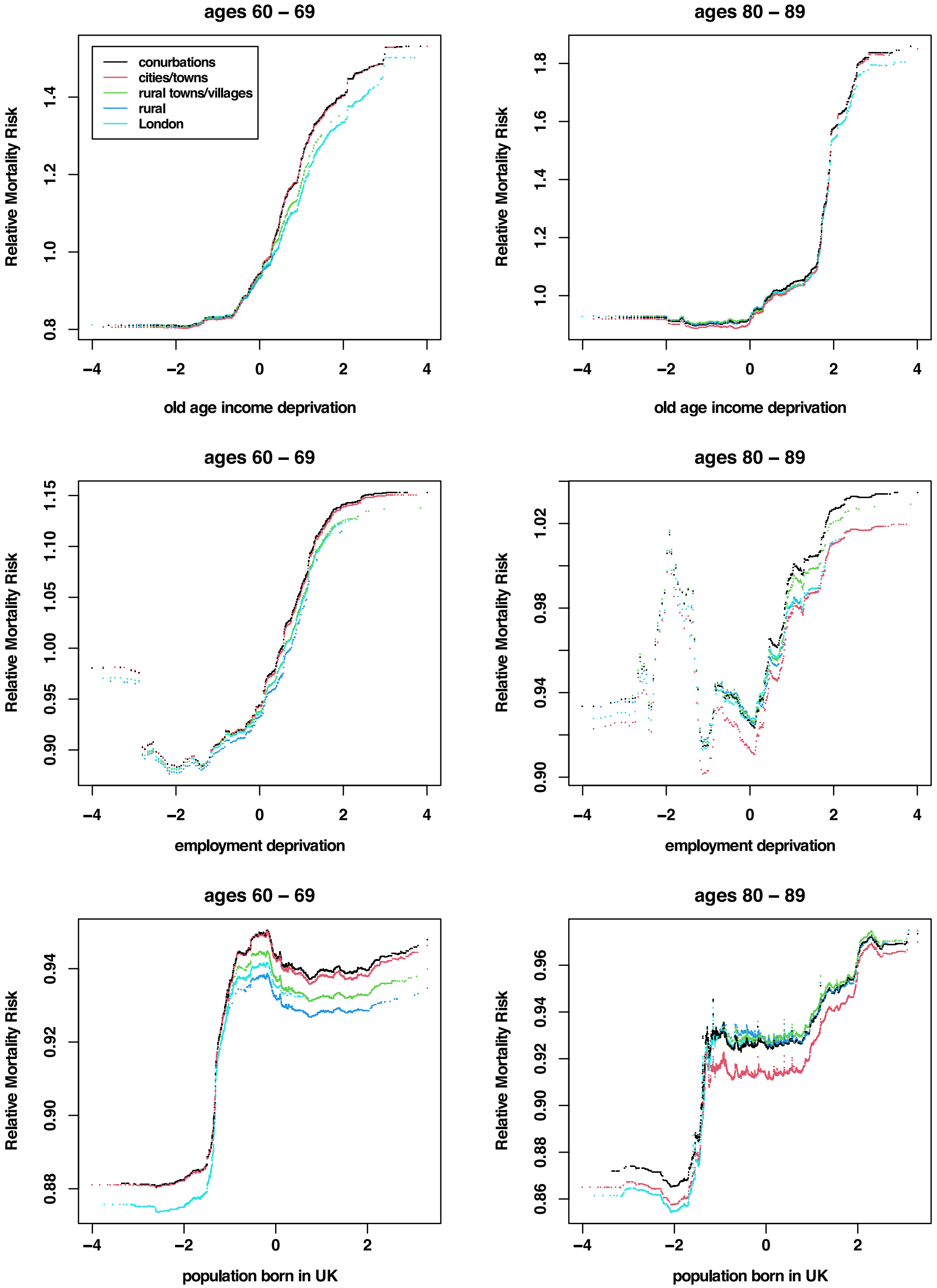

Figure 8. Estimated relative risk as a function of income deprivation at old age (first row), employment deprivation (second row) and the proportion of the population born in the UK (third row). In each case, other covariates (except the urban–rural classification) are fixed to their median across all LSOAs. The index is fitted to mortality data for ages 60–69 (left column) and ages 80–89 (right column), and urban–rural classes are colour coded.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

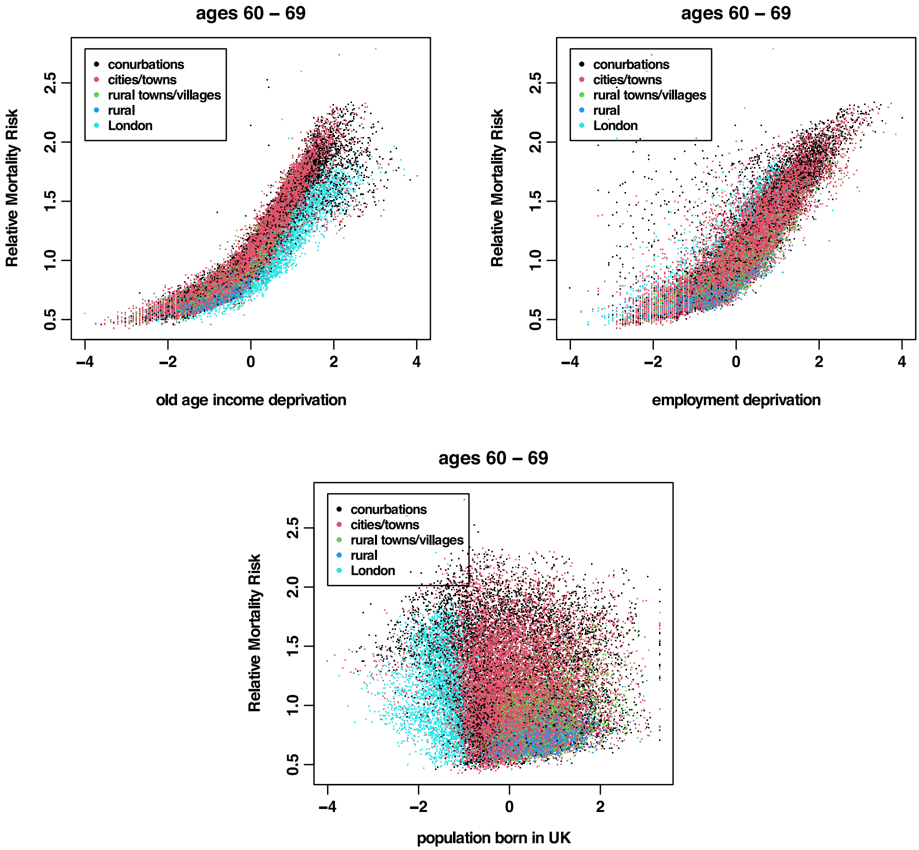

Figure 9. *LIFE index scores as a function of income deprivation at old age (top left), employment deprivation (top right) and the proportion of the population born in the UK (bottom). In contrast to Figure 8, all covariates other than care home populations, x_{11} and x_{12}, have been left at their observed values. The index is fitted to mortality data for ages 60–69, and urban–rural classes are colour coded.*

up, the divisions between colour bands gradually tilts indicating that there is more of a balance between the two measures of deprivation in terms of their impact on the LIFE index. This then gives a good indication of how the non-linear RF algorithm is easily able to pick up changes in the impact of different variables as we move across the dataset.

9. Deciles of the LIFE index

9.1. Relative risk in different deciles

With the proposed LIFE index, we can also zoom in on specific populations identified by their mortality risk. For example, Figure 11 shows the LIFE index values where we only consider the 10% of LSOAs with the lowest relative risk (left plot) and the 10% of LSOAs with the highest relative risk (right plot).

More specifically, we introduce a decile function g which identifies the decile k for any LSOA i with $g(i) = k$ for $k \in \{1, \ldots, 10\}$. The 10% of LSOAs with the highest mortality risk are LSOAs i with $g(i) = 1$ and the LSOAs with the lowest relative risk have $g(i) = 10$.

The plots in Figure 11 clearly show that in both subpopulations the most important variable to explain mortality differences is old-age income deprivation while employment deprivation has less impact.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

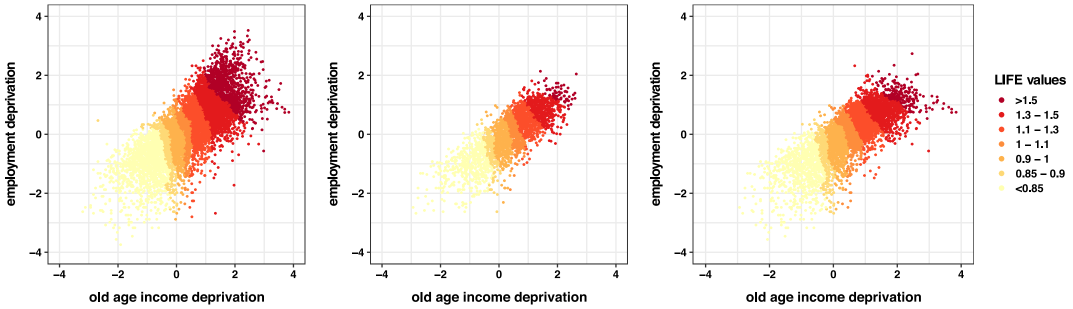

Figure 10. *Heatmaps of LIFE index values as a function of income deprivation at ages 65+ and employment deprivation. All other variables are fixed to the median. The left panel shows LSOAs in urban–rural class 1 (7921 LSOAs), the middle plot shows results for 2542 LSOAs in class 4 (isolated dwellings) and the right plot is for the 4810 LSOAs in London. The LIFE index is fitted to mortality data for ages 60–69.*

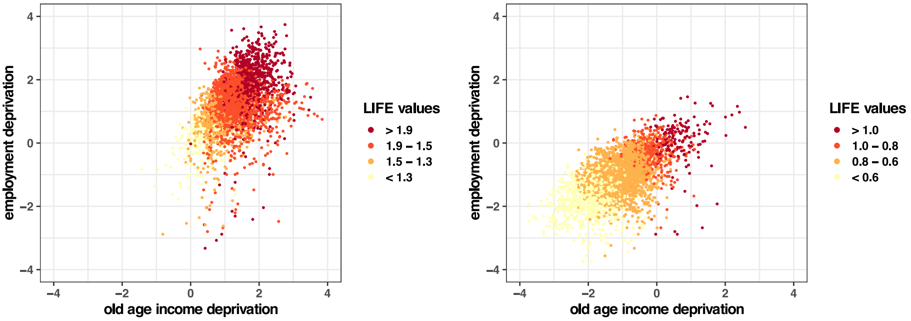

Figure 11. *Heatmaps of observed relative risk values as a function of income deprivation at ages 65+ and employment deprivation. All other variables are unchanged. The LIFE index is fitted to mortality data for ages 60–69. In the left plot, only LSOAs with $g(i) = 1$ (highest relative risk) are shown, and for the right plot, only LSOAs with $g(i) = 10$ (lowest) are used.*

9.2. Explanatory power of the LIFE index

The LIFE index has been constructed with the aim to explain the mortality risk in an LSOA based on socio-economic variables. Therefore, the question arises whether the LIFE index can indeed differentiate between high and low mortality areas. Or, in other words, are the socio-economic variables and the constructed index able to predict the relative mortality risk in an LSOA. To investigate this, we calculate age standardised mortality rates (ASMR) and ADSMR.

The ASMR for any population in year t is calculated as follows:

$$ASMR_{gt} = \frac{\sum_{a\in\mathcal{X}} m_{gta} E^s_a}{\sum_{a\in\mathcal{X}} E^s_a} \tag{9.1}$$

where $\mathcal{X}$ refers to the age range used and $m_{gta} = D_{gta}/E_{gta}$ is the crude death rate for the underlying population g in year t for age a. The standard population E^s_a used in our study is the European Standard Population[3] (ESP) in 2013.

To investigate how well the LIFE index can distinguish between different levels of mortality, we split the set of all LSOAs into 10 deciles as described above and plot the ASMRs for each decile over

[3] Report on ESP 2013 is available at: https://ec.europa.eu/eurostat/documents/3859598/5926869/KS-RA-13-028-EN.PDF/e713fa79-1add-44e8-b23d-5e8fa09b3f8f, downloaded 24 June 2019.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

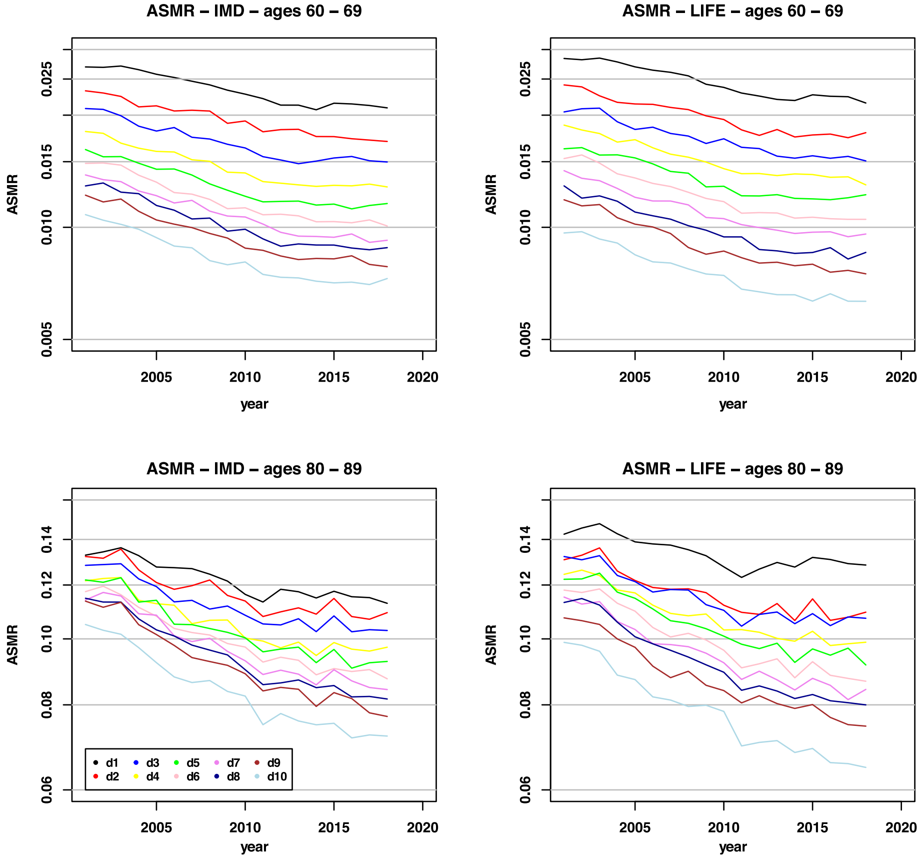

Figure 12. *ASMR by deprivation decile based on IMD scores (left) and LIFE scores (right) (log scale). The LIFE index and the ASMRs have been calculated for age groups 60–69 and 80–89. Similar figures for other age groups can be found in the supplementary material published online.*

time. We then compare our results to ASMRs calculated for deciles obtained from the Index of Multiple Deprivation (IMD), see Wen *et al*. (2021). The results are shown in Figure 12.

We can see in Figure 12 that the 10 deciles obtained from either index, LIFE or IMD, produce very different mortality rates. The figure also shows that the LIFE index leads to a wider spread of mortality rates meaning that it is better than the IMD in identifying low and high mortality on the basis of socio-economic variables. However, we must keep in mind that the IMD was not designed to predict mortality while the LIFE index was chosen to do so. We also find that both indices show a widening of the mortality gap, the difference between rates for the most deprived as compared to the least deprived. For the IMD, we discussed this issue in detail in Wen *et al*. (2021).

9.3. Explaining regional differences

While the LIFE index only uses information about socio-economic covariates, including an urban–rural class, to predict the relative mortality risk in an LSOA, it might be that LSOAs in different regions in England have different mortality rates although they have the same socio-economic characteristics.

To investigate this question further, we group all LSOAs into nine geographical regions. Table 8 lists the regions and shows the percentage of a region's LSOAs that belong to the risk groups defined in Section 7 for the LIFE index fitted to ages 60–69.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

Table 8. Distribution of LSOA into regions for different subpopulation classes. The LIFE index has been fitted to ages 60–69.

Region	All LSOAs	$G^l_{0.05}$	$G^u_{0.05}$	$G^l_{0.5}$	$G^u_{0.5}$
East	3614	197 (5.4)	59 (1.6)	2198 (60.8)	1416 (39.2)
East Midlands	2774	69 (2.5)	103 (3.7)	1442 (52.0)	1332 (48.0)
London	4835	205 (4.2)	81 (1.7)	1895 (39.2)	2940 (60.8)
North East	1657	35 (2.1)	180 (10.9)	557 (33.6)	1100 (66.4)
North West	4497	114 (2.5)	556 (12.4)	1832 (40.7)	2665 (59.3)
South East	5382	616 (11.4)	76 (1.4)	3531 (65.6)	1851 (34.4)
South West	3281	179 (5.5)	75 (2.3)	2040 (62.2)	1241 (37.8)
West Midlands	3487	107 (3.1)	234 (6.7)	1482 (42.5)	2005 (57.5)
Yorkshire	3317	121 (3.6)	279 (8.4)	1445 (43.6)	1872 (56.4)
Total	32,844	1643 (5)	1643 (5)	16,422 (50)	16,422 (50)

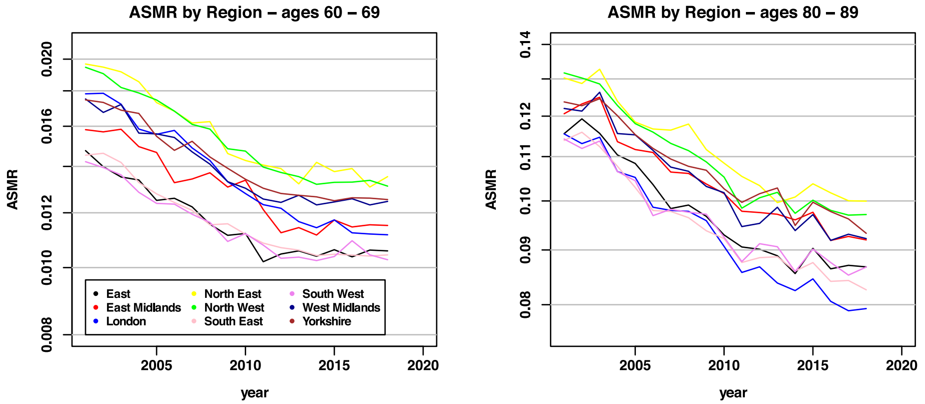

Figure 13. ASMRs by region for mortality data for ages 60–69 and 80–89 (log scale). Similar figures for other age groups can be found in the supplementary material published online.

The table shows that 12.4% of the LSOAs in the North West of England belong to the group of LSOAs with the 5% highest mortality risk, closely followed by the North East. The lowest LIFE scores are observed in the South East with 11.4% of those LSOAs belonging to the 5% English LSOAs with the lowest mortality risk. The results for other age groups are very similar, and, therefore, not reported in this paper.

To obtain a more detailed picture of the regional differences, we consider age standardised mortality rates. The ASMRs by region for age 60–69 and 80–89 can be seen in the Figure 13. We find that there are apparently substantial inequalities between regions, but how much of that can be explained by differences in the socio-economic mix of the nine regions?

To address this question, we propose using what we call the Age and Deprivation Standardised Mortality Rate, ADSMR. In the same way that the basic ASMR removes differences between populations that have different age profiles, the ADSMR is designed to remove the impact of different deprivation profiles. Thus, the ADSMR in region r in year t is defined as

$$ADSMR_{rt} = \frac{1}{10} \sum_{k=1}^{10} ASMR_{rkt} \qquad (9.2)$$

where $ASMR_{rkt}$ is the ASMR in year t of all LSOAs in region r and index decile k. Deciles are formed on the basis of the LIFE index values as described above, and, for comparison, on an IMD basis.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

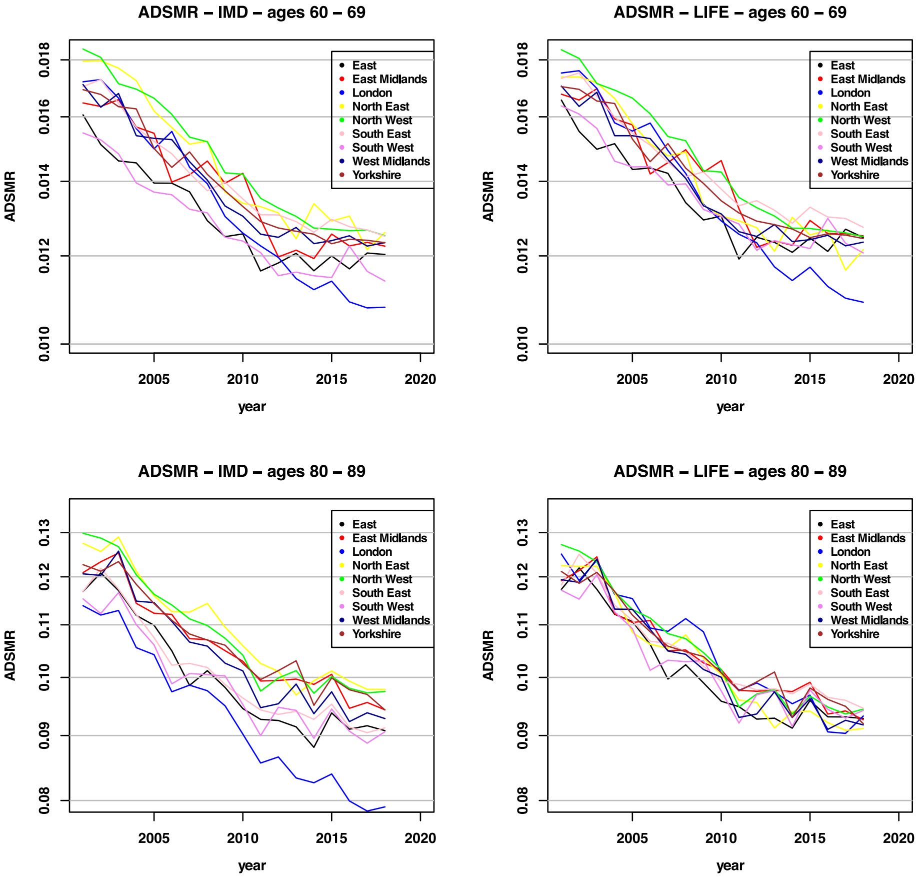

Figure 14. *ADSMRs on the basis of IMD deciles (left) and LIFE deciles (right) where mortality data for different age groups have been used (log scale). Similar figures for other age groups can be found in the supplementary material published online.*

Assuming that all mortality differences are explained by the LIFE (or IMD) index based on socio-economic variables, the ADSMRs should be the same for different regions (since $ASMR_{rkt}$ would be independent of r). In Figure 14, we plot $ADSMR_{rt}$ for the nine regions where deciles have been obtained from using the IMD (left plots) and the LIFE index (right plots).

We clearly see that the purpose built LIFE index is better suited to explain mortality differences between regions, resulting in much smaller differences in the ADSMR between the regions than the IMD-based ADSMR. While this is true for both age ranges (and other ranges as shown in Figures published online in the supplementary material), it is for the oldest age group 80–89 that the LIFE index can explain most of the differences between regions – the ADSMRs in the bottom right plot are much more similar than in the bottom left plot. This can also be observed for age group 70–79, see the online supplementary material for the relevant graphs. The reason for the much better ability of the LIFE index to explain differences between regions for higher age groups as compared to the IMD might be the inclusion of old-age income deprivation in the LIFE index rather than general income deprivation (which is one of the domains of deprivation used for the IMD).

Another interesting feature we observe in Figure 14 and similar figures in the online supplementary material are the mortality improvement rates in London which seem to be much greater than in other regions. Considering the oldest age group 80–89 and measuring deprivation using the IMD suggests that mortality in London improved significantly more than in other areas even after accounting for deprivation and that this improvement continued after 2011 when other regions have experienced no or little improvements. However, measuring deprivation using the LIFE index changes this conclusion.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press

The ADSMRs for the nine regions are very close together, and there is no London-effect visible for this age range.

10. Summary and conclusion

We have introduced a new mortality index, the LIFE that uses socio-economic characteristics to explain mortality rates in individual LSOAs. The LIFE index is constructed by first modelling the relationship between mortality and explanatory variables as a non-parametric function and estimating that function using the RF method. In a second step, the resulting regression function is adjusted to account for inflated death counts in LSOAs with care homes so that the LIFE index is a good representation of the general population.

Using the RF estimator of the relationship between socio-economic variables and relative mortality risk, we are able to study the impact of specific variables in isolation. While we have only reported results for three covariates in this paper, the proposed method allows for a much more detailed analysis. However, our empirical results indicate that of all the covariates considered it is old-age income deprivation which has the highest explanatory power for mortality differences between LSOAs – the higher old-age income deprivation the higher are the mortality rates in an LSOA.

While studying the impact of individual variables helps to understand to some extent how the LIFE index allocates scores to LSOAs with certain characteristics, the impact of individual factors is still not as clear as it is in parametric models. A careful analysis of the sensitivity of the LIFE index scores with respect to changes in certain factors could serve as a first step in the development of a parametric (or semiparametric) regression model in which estimated parameters have a clear interpretation.

Comparing the LIFE index with the widely used English Index of Multiple Deprivation shows that the LIFE index is better able to explain regional variations in mortality with deprivation measures than the IMD. However, keeping in mind that the IMD has not been constructed with reference to relative mortality risk, we find that it actually is a good predictor for mortality differences. Nevertheless, for any application with the goal of explaining mortality differences between different geographic locations in England the LIFE index is more suitable than the IMD.

The proposed LIFE index could be further improved in different directions. Other covariates could, of course, be considered. One LSOA-specific covariate that might be important is the average age in an LSOA within a given age band (e.g. 60–69). The construction of the relative risk takes the age structure of LSOAs into account (both expected and observed deaths are age dependent). However, it might be that certain covariates have different effects in LSOAs with higher or lower average age. This is different from considering different age ranges for fitting the relative risk to observed covariates.

Another extension of our index would be to make it time dependent. However, that would require the measurement of the LSOA-specific relative risk for individual calendar years, which would increase the variance of the observed relative risk substantially.

Acknowledgements. This work forms part of the research programme "Modelling, Measurement and Management of Longevity and Morbidity Risk" funded by the Actuarial Research Centre of the Institute and Faculty of Actuaries, the Society of Actuaries and the Canadian Institute of Actuaries. The authors gratefully acknowledge feedback from the respective steering and oversight groups.

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.

Supplementary Material. To view supplementary material for this article, please visit http://doi.org/10.1017/asb.2023.12.

References

Bennett, J.E., Li, G., Foreman, K., Best, N., Kontis, V., Pearson, C., Hambly, P. and Ezzati, M. (2015) The future of life expectancy and life expectancy inequalities in England and Wales: Bayesian spatiotemporal forecasting. *Lancet*, **386**(9989), 163–170.

Bjerre, D.S. (2022) Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin: The Journal of the IAA*, **52**(3), 765–787.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**(1), 5–32.

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., Kleinow, T. and Wen, J. (2021) Drivers of mortality: Risk factors and inequality. Working Paper, Heriot-Watt University. http://www.macs.hw.ac.uk/ndrewc/ARCresources/LIFEapp/LocalLinearRegression20210607.pdf

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. *JAMA*, **315**(16), 1750–1766.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(1), 3.

Hong, W.H., Yap, J.H., Selvachandran, G., Thong, P.H. and Son, L.H. (2021) Forecasting mortality rates using hybrid lee–carter model, artificial neural network and random forest. *Complex & Intelligent Systems*, **7**(1), 163–189.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning*. New York, Heidelberg, Dordrecht, London: Springer Press.

Levantesi, S. and Nigri, A. (2020) A random forest algorithm to improve the lee–carter mortality forecasting: Impact on q-forward. *Soft Computing*, **24**(12), 8553–8567.

Levantesi, S. and Pizzorusso, V. (2019) Application of machine learning to mortality modeling and forecasting. *Risks*, **7**(1), 26.

Liaw, A. and Wiener, M. (2002) Classification and regression by RandomForest. *R News*, **2**(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

Lyu, P., Li, J.S.-H. and Zhou, K.Q. (2022) Socioeconomic differentials in mortality: Implications on index-based longevity hedges. *Scandinavian Actuarial Journal*, 1–29. https://doi.org/10.1080/03461238.2022.2104131

Mackenbach, J.P., Bos, V., Andersen, O., Cardano, M., Costa, G., Harding, S., Reid, A., Hemström, O., Valkonen, T. and Kunst, A.E. (2003) Widening socioeconomic inequalities in mortality in six Western European countries. *International Journal of Epidemiology*, **32**(5), 830–837.

Mackenbach, J.P., Kulhánová, I., Menvielle, G., Bopp, M., Borrell, C., Costa, G., Deboosere, P., Esnaola, S., Kalediene, R., Kovacs, K., Leinsalu, M., Martikainen, P., Regidor, E., Rodriguez-Sanz, M., Strand, B.H., Hoffmann, R., Eikemo, T.A., Östergren, O., Lundberg, O. and Eurothine and EURO-GBD-SE consortiums (2015) Trends in inequalities in premature mortality: A study of 3.2 million deaths in 13 European countries. *Journal of Epidemiology and Community Health*, **69**(3), 207–17; discussion 205–6.

Mayhew, L., Harper, G. and Villegas, A.M. (2020) An investigation into the impact of deprivation on demographic inequalities in adults. *Annals of Actuarial Science*, **14**(2), 358–383.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raleigh, V.S. and Kiri, V.A. (1997) Life expectancy in england: Variations and trends by gender, health authority, and level of deprivation. *Journal of Epidemiology & Community Health*, **51**(6), 649–658. https://jech.bmj.com/content/51/6/649

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D. and Plunkett, E. (2015) The English indices of deprivation 2015. Research Report, Department for Communities and Local Government.

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**(1), 168–193.

Wen, J. (2022) *Multi-population and factor-based mortality analytics*. Ph.D. Thesis, Heriot-Watt University, Edinburgh. http://hdl.handle.net/10399/4554

Wen, J., Cairns, A.J.G. and Kleinow, T. (2021) Fitting multi-population mortality models to socio-economic groups. *Annals of Actuarial Science*, **15**(1), 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.

Woods, L.M., Rachet, B., Riga, M., Stone, N., Shah, A. and Coleman, M.P. (2005) Geographical variation in life expectancy at birth in england and wales is largely explained by deprivation. *Journal of Epidemiology & Community Health*, **59**(2), 115–120.

https://doi.org/10.1017/asb.2023.12 Published online by Cambridge University Press