A New Socio-Economic Mortality Index for England

Jie Wen
joint work with Andrew J.G. Cairns and Torsten Kleinow

Heriot-Watt University, Edinburgh

School of Mathematical and Computer Sciences

Sixteenth International Longevity Risk and Capital Markets Solutions Conference

August 2021

Content

Content

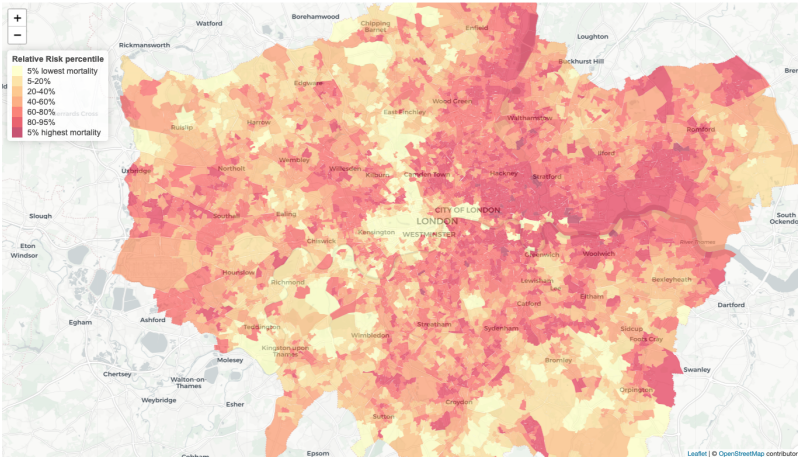
Overview

The Longevity Index For England ("LIFE") Index:

- England contains $N = 32,844$ geographical units called Lower-layer Super Output Areas (LSOAs) – each of them has population size of around $1,500$.
- Population in one LSOA in general have socio-economic homogeneity.
- The index measures the average mortality in each LSOA relative to the national average mortality by gender and age, and covers mortality experience over a selected period (2001-2018).
- It is constructed using a random forest algorithm that takes relevant socio-economic variables as input.
- Distribution of the index is centred at 1 (close to the national average).

Overview (cont.)

Example: LSOAs in London and coloured by the LIFE Index value for males of age 75 (Source: LIFE App):



Content

Data and Modelling Framework

Neighbourhood-level mortality data in England are used for constructing the index:

- Gender-specific death and exposure counts in individual LSOAs – D_{itx} and E_{itx} available for every singe LSOA i, year t and age x.
- We focus on population of "pensionable ages" – over ages 60-89.
- **Predictive variables**: socio-economic factors at LSOA-level, denoted as $X_{i,j}$ (the j^{th} variable in LSOA i):
  - There are numerical and categorical variables.
  - Most are gender neutral and homogeneous over all ages.
  - Not updated frequently over time.
- **Response variable**: relative mortality risk, R_i^0, in single LSOAs i by single age, derived from observed D_{itx} and E_{itx} using rolling 10-year age intervals, i.e. data of age 60-69 for age 65. Having $m_{tx}^b = \sum_i D_{itx} / \sum_i E_{itx}$ as the national average mortality by single year and age:

$$\mathbf{R}_i^0 = \frac{D_i}{\hat{D}_i^0} = \frac{\sum_{tx} D_{itx}}{\sum_{tx} E_{itx} m_{tx}^b}$$

Data and Modelling Framework (cont.)

LSOA-level **predictive variables** relate to socio-economics:

X_1	old-age income deprivation
X_2	employment deprivation (i.e. unemployment)
X_3	education deprivation
X_4	housing standard (number of bedrooms)
X_5	proportion of the population born in the UK
X_6	deprivation in housing/living environment
X_7	employment/occupation: proportion in a management position
X_8	crime rate
X_9	proportion working more than 49h per week
X_{10}	proportion of population aged 60+ in a care home with nursing
X_{11}	proportion of population aged 60+ in a care home without nursing
X_{12}	urban-rural class (value of 1 to 5)

$\mathbf{X} = (X_1, \ldots, X_{12})$ represents the $32,844 \times 12$ variable space.
$\mathbf{x}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,12})$ is the known socio-economic features of one particular LSOA i.

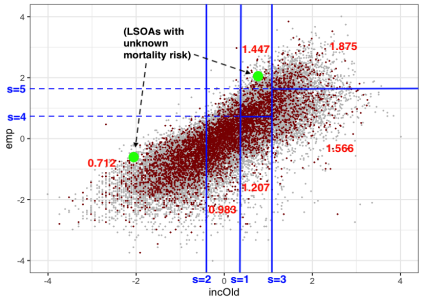
Content

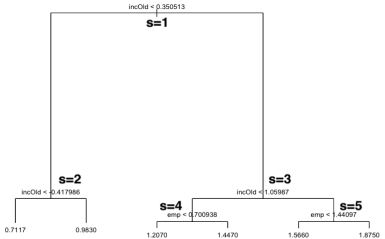
Regression tree:

- There are B regression trees within the random forest ("RF") algorithm, and are denoted as $\hat{f}^{(b)}$, $b \in (1, 2, \ldots, B)$.
- Every $\hat{f}^{(b)}$ is trained using a randomly selected subset of underlying LSOAs rather than all of them, i.e. they are not trained using exactly the same observations.
- $\hat{f}^{(b)}$ is derived using binary splits made to the dataset in reference to the $\mathbf{X}$. It stratifies all underlying LSOAs into disjoint groups called nodes.
- All LSOAs within one node have the same estimate by $\hat{f}^{(b)}$, as the average of all observed R^0's in this node.
- $\hat{f}^{(b)}$ is a piecewise constant function.

Regression Tree and Random Forest (cont.)

Stylized example: A single regression tree model $\hat{f}^{(b)}$ trained using the observed R^0 and two predictive variables – old-age income deprivation score ($X_1 = incOld$) and employment deprivation score ($X_2 = emp$).





For example: $\hat{f}^{(b)}(\mathbf{x}) = 1.207 \quad \forall \mathbf{x} \in \{\mathbf{x} : 0.351 \leq X_1 < 1.060 \quad \text{and} \quad X_2 < 0.701\}$.

(*) This example only has 6 nodes – in real-world application there are always much more nodes, i.e. we use 30 to 50 nodes per tree for constructing the index.

Regression Tree and Random Forest (cont.)

Random forest – ensemble of regression trees:

- In every single $\hat{f}^{(b)}$, only a randomly selected 4 out of 12 predictive variables are considered while making every split.
- Estimation from the RF, $\hat{f}^{RF}$, is the average over all individual trees' estimation:

$$\hat{f}^{RF}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{(b)}(\mathbf{x})$$

- Both sample and variable randomness are introduced into the RF algorithm and mitigates overfitting risk in single regression trees.

Content

We applied the random forest to construct a LIFE Index, which is constructed based on the estimate of relative risk in single LSOAs, $\hat{f}^{RF}(\mathbf{x})$, and with adjustment made for impact of presence of care homes on the mortality level in one LSOA:

- Instead of using the actual observed variables of care home, we use the national average proportion of population (age 60+) in care homes as constant variables in all LSOAs.

$$I_i = \hat{f}^{RF}(\tilde{\mathbf{x}}_i)$$

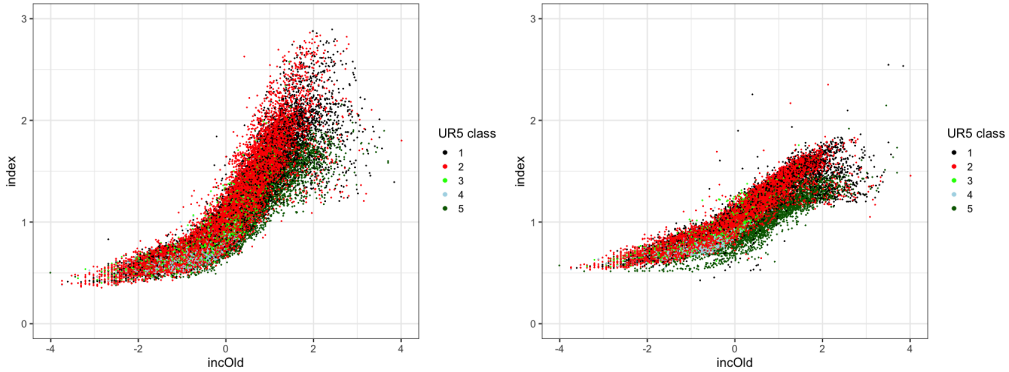
$\tilde{\mathbf{x}}_i$ is the socio-economic features in one LSOA i, with the two care home variables, $X_{i,10}$ and $X_{i,11}$ replaced by their mean average value over all $32,844$ LSOAs in England, i.e.

$$\tilde{\mathbf{x}}_i = \left(X_{i,1}, X_{i,2}, \ldots, \bar{X}_{10}, \bar{X}_{11}, X_{i,12}\right)$$

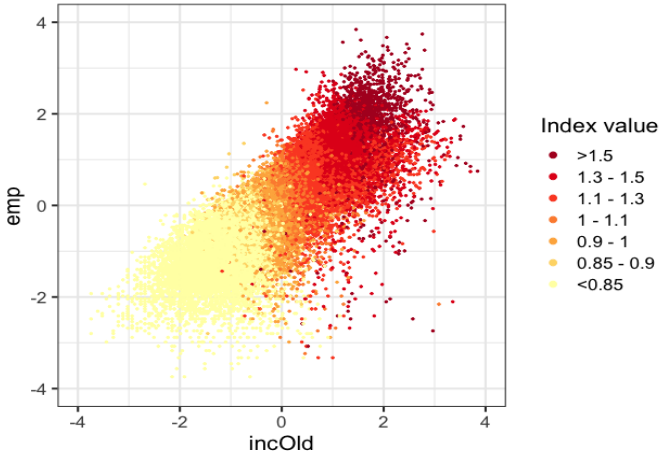
for any one LSOA i.

Results (cont.)

Index value of England males population for age 65 (left) and 75 (right), with years 2001-2018 taken into account, plotted over *incOld* as one of the predictive variables:

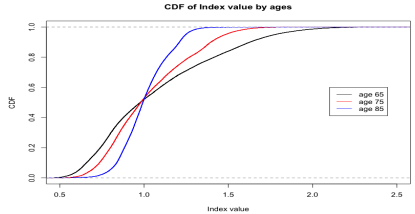


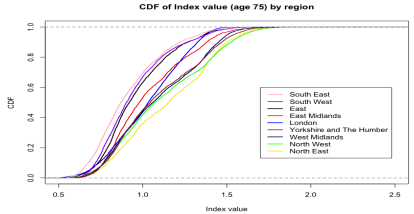
LIFE Index of England males of age 75, in heat plot over two predictive variables:

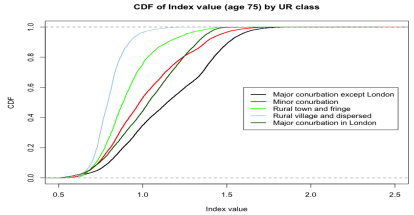


Results (cont.)

Cumulative distribution function (CDF) plots showing distribution of LSOAs grouped in different ways:







Content

Summary

- The LIFE Index explains mortality difference over different areas in England by relevant socio-economic factors, and is available by age and gender.
- We are in-progress of a working paper about using random forest algorithm to create the mortality index.
- There is an online App for non-expert users to explore the LIFE Index and discover mortality inequalities between different areas of England.
  tinyurl.com/LIFEindex
- There is also a webinar that covers more technical details about the LIFE Index. See the IFoA ARC website for more details.
  https://www.actuaries.org.uk/learn-and-develop/research-and-knowledge/actuarial-research-centre-arc

THANK YOU! & Any Questions?

wenjiese7en@gmail.com

HERIOT WATT UNIVERSITY

Actuarial Research Centre
Institute and Faculty of Actuaries