Modelling Neighbourhood Mortality Using the Random Forest

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

Heriot-Watt University, Edinburgh

School of Mathematical and Computer Sciences

June 2021

The views expressed in this presentation are those of invited contributors and not necessarily those of the Institute and Faculty of Actuaries (IFoA).

The IFoA does not endorse any of the views stated, nor any claims or representations made in this presentation and accept no responsibility or liability to any person for loss or damage suffered as a consequence of theirplacing reliance upon any view, claim or representation made in this presentation.

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

Content

1. Data and Modelling Framework
2. Regression Tree and Random Forest
3. Results
4. Summary

Content

Data and Modelling Framework

Neighbourhood-level mortality data in England:

- England contains $N = 32,844$ small geographical areas (neighbourhoods) called Lower-layer Super Output Areas (LSOA) – each of them has population size of around $1,500$. Population in one LSOA in general have socio-economic homogeneity.
- 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" are what we focus on, i.e. ages 60-89.
- **Predictive variables**: socio-economic factors are available at LSOA-level, denoted as X_{ij} (the j^{th} variable in LSOA i). There are numerical metrics and categorical metrics. They are gender neutral and homogeneous over all ages modelled.
- **Response variables**: relative mortality risks in every LSOA by single age, derived using rolling 10-year age intervals, i.e. data of age 60-69 to represent age 65, 70-79 for 75 80-89 for 85, etc.

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 inside 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 classification

$\mathbf{X} = (X_1, \ldots, X_{12})$ represents a $32,844 \times 12$ variable space.

Data and Modelling Framework (cont.)

Five levels of urbanization (X_{12}) are defined for every single LSOA:

- Class 1 ($X_{12} = 1$): Urban conurbation (except London)
- Class 2 ($X_{12} = 2$): Urban city and town
- Class 3 ($X_{12} = 3$): Rural town and fringe
- Class 4 ($X_{12} = 4$): Rural village and dispersed
- Class 5 ($X_{12} = 5$): Urban conurbation in London

Data and Modelling Framework (cont.)

LSOA-level **response variable** is relative mortality risk. It is defined as an actual-vs-expected death ratio ("A-E ratio") in every LSOA i:

$$m^b_{tx} = \frac{\sum_i D_{itx}}{\sum_i E_{itx}} \quad \text{national average death rate by single ages and years}$$

$$D_i = \sum_{tx} D_{itx} \quad \text{actual aggregate number of deaths over selected ages and years, by single LSOAs}$$

$$\hat{D}^0_i = \sum_{tx} m^b_{tx} E_{itx} \quad \text{"expected" aggregate deaths over selected ages and years, by single LSOAs}$$

$$R^0_i = \frac{D_i}{\hat{D}^0_i} \quad \text{A-E ratio over selected ages and years, by single LSOAs}$$

R^0_i measures the empirical relative mortality risk in one LSOA i relative to national average mortality. It can be calculated over narrow age groups to capture differenct trends of mortality by age.

Data and Modelling Framework (cont.)

We train the random forest model using the $N = 32,844$ LSOAs, by split them into two halves as disjoint subsets:

- The training set, $\mathcal{S}^{tr} \subset \{1, 2, \ldots, N\}$, contains LSOAs used in training the random forest model.
- The validation set, $\mathcal{S}^{va} \subset \{1, 2, \ldots, N\}$, contains LSOAs used for model validation. They are not directly involved in model training and only used for selecting the hyperparameters.
- The two sets are disjoint, i.e. $\mathcal{S}^{tr} \cap \mathcal{S}^{va} = \emptyset$, and together form the full sample of N LSOAs, i.e. $\mathcal{S}^{tr} \cup \mathcal{S}^{va} = \{1, 2, \ldots, N\}$.

Hyperparameters are parameters in non-parametric models that control the bias-variance trade-off, i.e. balance between underfitting and overfitting.

Content

We firstly look at single regression trees – random forest ("RF") model is an ensemble of multiple regression trees.

The observed actual-versus-expected ratios ("A-E ratio"), $y = (R_1^0, \ldots, R_N^0)$, along with socio-economic factors $\mathbf{X} = (X_1, \ldots, X_{12})$ are the data used to train the regression trees.

$\hat{f}^{(b)}$ with $b \in (1, 2, \ldots, B)$ is one of the regression tree functions. It is non-parametric and do not have any closed-form formula.

Every single $\hat{f}^{(b)}$ in RF model is trained using a randomly selected subset of LSOAs ($\mathcal{S}^b$) of the training set, i.e. $\mathcal{S}^b \subset \mathcal{S}^{tr}$.

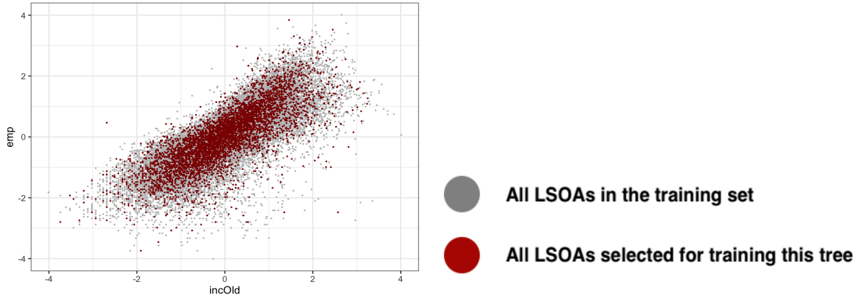
$\hat{f}^{(b)}(\mathbf{x}_i)$ is the estimate of relative risk in LSOA i using the known socio-economic factors in $\mathbf{x}_i$, i.e. $\mathbf{x}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,12})$.

They way in which every single tree $\hat{f}^{(b)}$ is trained:

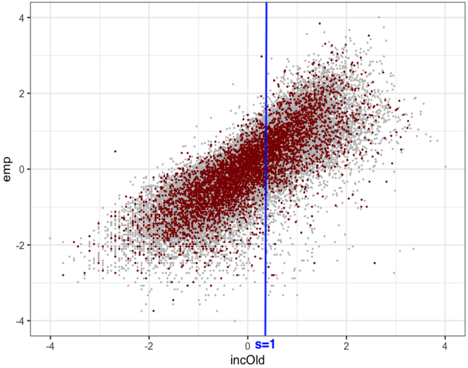
- $\hat{f}^{(b)}$ is derived following binary splits in reference to $\mathbf{X}$, which stratify the training set is stratified into disjoint groups of LSOAs called nodes.
- All the splits are made iteratively to the existing nodes created by earlier splits.
- Three factors to consider for every split – which existing node to split, by which predictive variable, and by what value as the split boundary.
- In general, the principle is to optimize improvement in accuracy of the model by making every split.
- All LSOAs within the same node have the same estimate by the $\hat{f}^{(b)}$, which is the average observed R^0 over this node.
- $\hat{f}^{(b)}$ is a piecewise constant function.

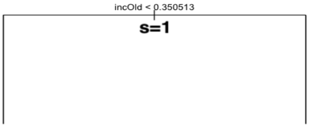
Stylized example: A single regression tree model $\hat{f}^{(b)}$ trained using the observed R^0 and two predictive variables – old-age income deprivation score ($X_1 = incOld$) and employment deprivation score ($X_2 = emp$).

Every dot represents one of the LSOAs – we know their observed R^0 and value of the two predictive variables applied:

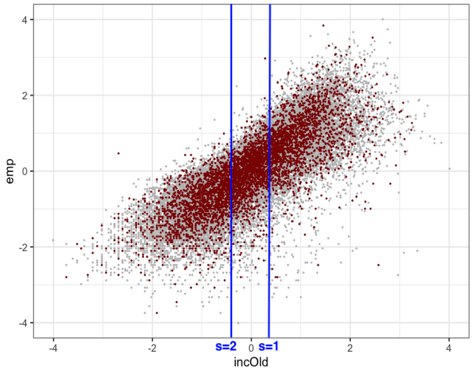


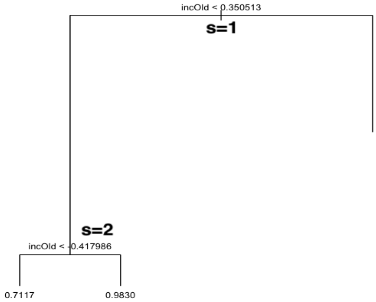
The first split made to the variable space:



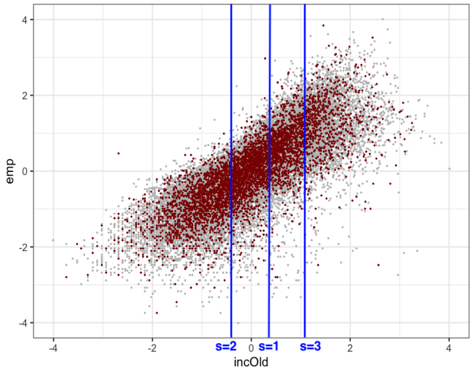


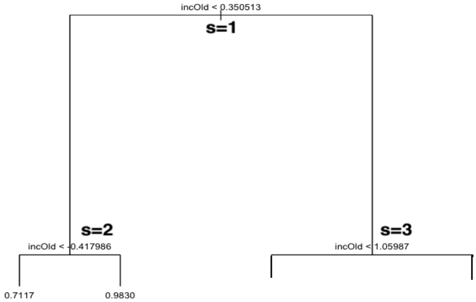
The second split made to the variable space:



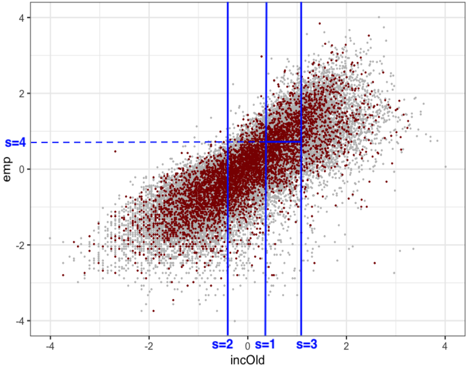


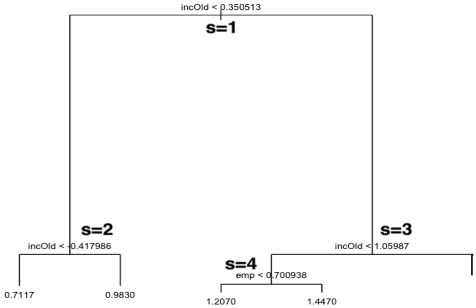
The third split made to the variable space:



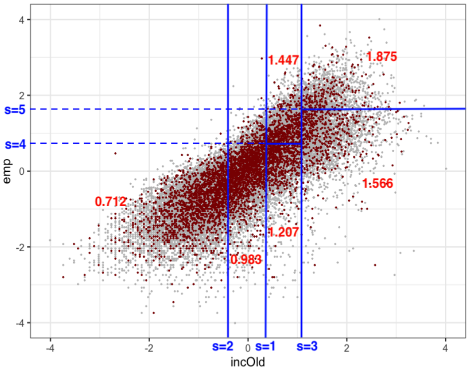


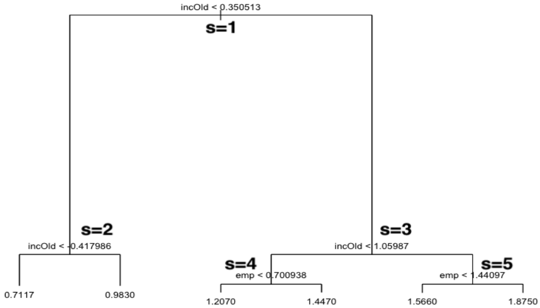
The forth split made to the variable space:



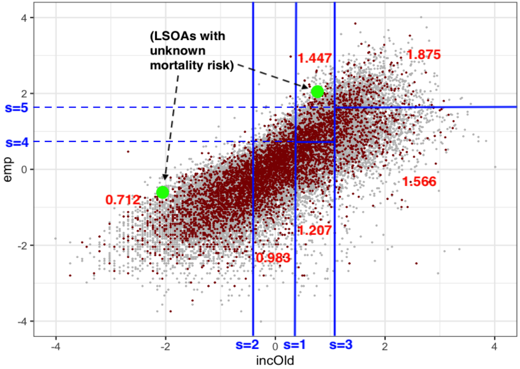


The fifth split (last one) made to the variable space:





The tree $\hat{f}^{(b)}$ can be used to predict relative risk in 'unknown' LSOAs using socio-economics input. Note that this is a simple example that only has 6 nodes – in the random forest model we get eventually there are around 30 to 50 nodes per tree.



The stylized example $\hat{f}^{(b)}$ can also be written as a piecewise constant function:

$$\hat{f}^{(b)}(\mathbf{x}) = \begin{cases} 0.712, & \forall \mathbf{x} \in \{\mathbf{x} : X_1 < -0.418\} \\ 0.983, & \forall \mathbf{x} \in \{\mathbf{x} : -0.418 \leq X_1 < 0.351\} \\ 1.207, & \forall \mathbf{x} \in \{\mathbf{x} : 0.351 \leq X_1 < 1.060 \quad \text{and} \quad X_2 < 0.701\} \\ 1.447, & \forall \mathbf{x} \in \{\mathbf{x} : 0.351 \leq X_1 < 1.060 \quad \text{and} \quad X_2 \geq 0.701\} \\ 1.566, & \forall \mathbf{x} \in \{\mathbf{x} : X_1 \geq 1.060 \quad \text{and} \quad X_2 < 1.441\} \\ 1.875, & \forall \mathbf{x} \in \{\mathbf{x} : X_1 \geq 1.060 \quad \text{and} \quad X_2 \geq 1.441\} \end{cases}$$

When to stop making splits in one tree:

- Different stopping criteria applicable. One we apply is a constraint that every node should have at least $M = 200$ LSOAs, otherwise no futher split is allowed.
- The existing nodes after we stop making further splits are called terminal nodes. They define the estimate of relative risk produced by the tree $\hat{f}^{(b)}$.

Variable randomness:

- In every single tree within RF model, only a randomly selected 4 out of 12 predictive variables are considered while making every split.

LSOAs and predictive variables considered in one particular split s of one particular tree b in a RF model (N^{tr} is the volume of training set used to train the RF model, and there are $p = 12$ predictive variables):

$$\mathbf{X}^{tr} = \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} & \cdots & X_{1,p-1} & X_{1p} \\ X_{21} & X_{22} & X_{23} & X_{24} & \cdots & X_{2,p-1} & X_{2p} \\ X_{31} & X_{32} & X_{33} & X_{34} & \cdots & X_{3,p-1} & X_{3p} \\ X_{41} & X_{42} & X_{43} & X_{44} & \cdots & X_{4,p-1} & X_{4p} \\ X_{51} & X_{52} & X_{53} & X_{54} & \cdots & X_{5,p-1} & X_{5p} \\ X_{61} & X_{62} & X_{63} & X_{64} & \cdots & X_{6,p-1} & X_{6p} \\ X_{71} & X_{72} & X_{73} & X_{74} & \cdots & X_{7,p-1} & X_{7p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \\ X_{N^{tr}-1,1} & X_{N^{tr}-1,2} & X_{N^{tr}-1,3} & X_{N^{tr}-1,4} & \cdots & X_{N^{tr}-1,p-1} & X_{N^{tr}-1,p} \\ \boxed{X_{N^{tr}1}} & X_{N^{tr}2} & \boxed{X_{N^{tr}3}} & \boxed{X_{N^{tr}4}} & \cdots & \boxed{X_{N^{tr},p-1}} & X_{N^{tr}p} \end{pmatrix}$$

Regression Tree and Random Forest (cont.)

The trained RF model denoted as $\hat{f}^{RF}$ is an ensemble of B regression trees, $\hat{f}^{(b)}$ for $b \in \{1, 2, \ldots, B\}$. Estimate by the RF is the average over all individual trees' estimates:

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

Compared to single regression trees that are not robust to data, RF introduces both sample and variable randomnesses and therefore mitigates overfitting risk greatly.

Content

The final proposed RF model with for estimating relative mortality risk at LSOA level:

Parameter/Hyperparameter	Value
Number of trees	2,500
Total number of variables	12
Number of variables to consider per split	4
Minimum size of terminal nodes	200

About the parameters/hyperparameters:

- Number of trees selected to ensure the RF model achieves the maximum accuracy while keeping computational burden low.
- The twelve predictive variables are selected beforehand.
- Number of variables to consider in every split is selected using cross validation so that the out-of-sample MSE over the $\mathcal{S}^{va}$ is minimized.
- Minimum size of terminal nodes in all trees is selected so that we achieve a balance between variance over the individual trees and accuracy in estimation.

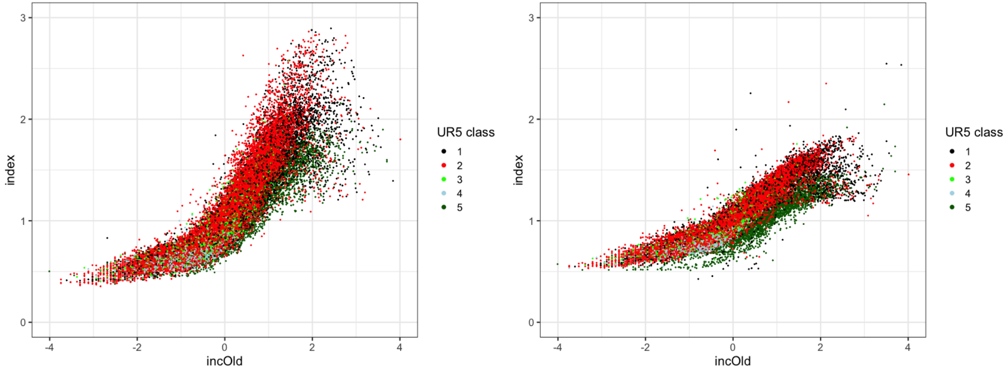
Results (cont.)

We applied the random forest to construct a Longevity Index for England ("LIFE") that measures the mortality level in one LSOA relative to the national level.

- It is created by gender, age and year (can be adjusted in the observed R^0).
- LSOAs having the index value close to 1 have mortality risk close to the national average (of certain gender and age/year group).
- It can be used as an indicator to mortality risk in one LSOA, or as a predictive variable alongside other factors to estimate mortality for individuals living in one LSOA, e.g. smoking status, long-term health condition, etc.

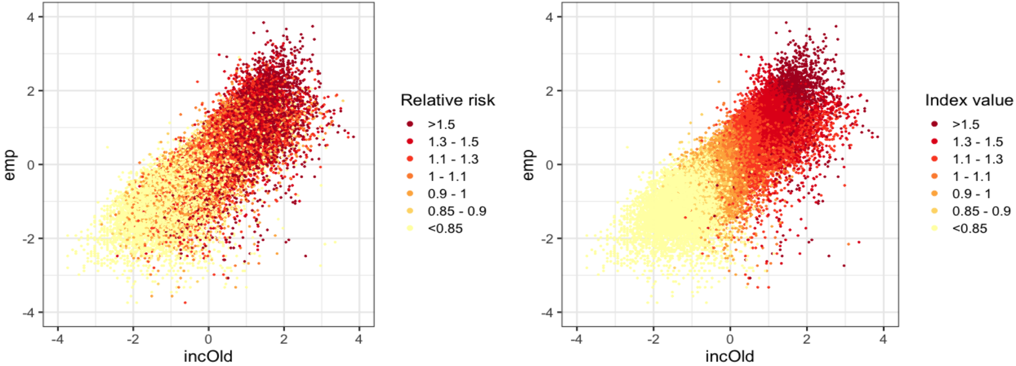
There is another ARC webinar presented by Andrew J.G. Cairns and Torsten Kleinow that has a complete discussion over the LIFE Index.

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:

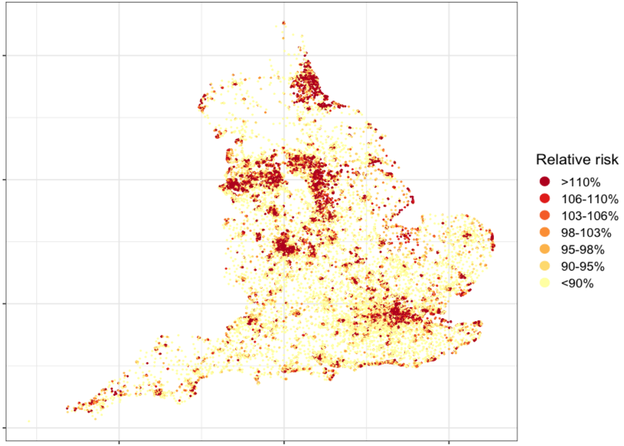


Results (cont.)

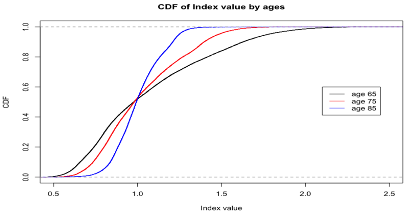
Comparison between observed R^0 (left) and LIFE Index (right) of England males population for age 75:

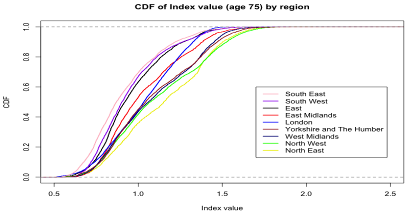


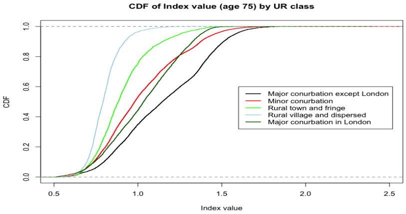
Distribution of the LIFE Index value for males population of age 75 in England:



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







Results (cont.)

Summary statistics of relative risk (males, age 75) by urban-rural class:

LSOAs	Min	1st Q.	Median	3rd Q.	Max
All	0.427	0.828	0.980	1.207	3.768
Urban conurbation except London (UR class 1)	0.427	0.926	1.145	1.387	3.768
Urban conurbation (UR class 2)	0.531	0.835	0.981	1.182	2.351
Rural town, fringe or village (UR class 3 and 4)	0.534	0.759	0.830	0.929	1.756
Urban conurbation in London (UR class 5)	0.512	0.861	1.034	1.206	2.146

According to distribution of the LIFE Index value:

- Mortality difference relevant to socio-economics is not significant in population of high ages.
- North of England in general have higher relative mortality risk than the South.
- LSOAs as conurbations and large cities (class 1) have the widest distribution of relative risks.
- Rural LSOAs (class 3 and 4) in general have lower relative risk than more urbanized ones.
- London (class 5) in general have lowest relative risk in all conurbations and large cities.

Content

1. Data and Modelling Framework
2. Regression Tree and Random Forest
3. Results
4. **Summary**

HERIOT WATT UNIVERSITY

Actuarial Research Centre
Institute and Faculty of Actuaries

Summary

Pros of random forest algorithm:

- Its non-parametric structure does not require prior assumption or knowledge about the functional relationship between response and predictive variables.
- It is invariant to transformation of the predictive variables.
- It captures potential interactions over the predictive variables, instead of needing us to set them up by experience or judgement.
- It runs faster than some other non-parametric models like local linear regression and kernel estimator.

However, it can be less interpretable than most parametric models for not having a closed-form formula.

HERIOT WATT UNIVERSITY

Actuarial Research Centre
Institute and Faculty of Actuaries

THANK YOU!

Any questions please: wenjiese7en@gmail.com / jw192@hw.ac.uk