Neighbourhood-level mortality modelling using socio-economic indicators

Andrey Ugarte Montero *[1], Ayse Arik †[2], and Andres Villegas Ramirez‡[3]

[1]Research Center for Longevity Risk, University of Amsterdam.
[2]School of Risk and Actuarial Studies, University of South Whales.
[3]School of Risk and Actuarial Studies, University of South Whales.

Abstract

In this study, we develop a modelling framework that quantifies the relationship between regional, demographic and socio-economic covariates through baseline mortality and age-related effects. This structure allows the influence of these components to be explained using spatial and socio-economic covariates that are fixed in time, such as those derived from census data that are not updated annually. Combining mortality modelling with machine learning techniques, the proposed framework isolates components of interest while allowing for a flexible class of models, including classical regression approaches, tree-based methods, boosting techniques, and generalized additive models, as well as hybrid combinations.

First, we estimate an empirical baseline mortality rate for each region. Then, we use a chosen learner to model this rate. We use a resampling technique similar to the Delete-a-Group Jackknife method to determine which covariates affect this baseline mortality. This technique computes the importance of each covariate available, while leaving one fold of the data out at a time. Such an analysis allows us to identify the relevant variables so we can create a baseline mortality model, and proceed to quantify the effect of aging in mortality. This aging effect is estimated with the help of residuals between the observed logmortality rate per region and age, and the estimation of the baseline mortality model. We first estimate a global aging effect using basis functions and Hermite splines. Next, we analyze how regions deviate from this pattern and focus on modeling these deviations using a learner that may differ from the one used for the baseline component. After performing orthogonalization of estimated effects with respect to the variables affecting baseline mortality, we select the covariates affecting mortality through aging, and construct an estimated lograte using the two components of the model. Model performance is assessed using simulated data under controlled settings, and we have confirmed that our approach satisfactorily captures the patterns in mortality introduced through the synthetic covariates in the simulation.

Therefore, the framework is applied to real data to model mortality at the Lower Super Output Area (LSOA) level in the UK between 2001 and 2022, using socio-economic and demographic

*E-mail address: a.d.ugartemontero@uva.nl
†E-mail address: a.arik@unsw.edu.au
‡E-mail address: a.villegas@unsw.edu.au

variables derived from census data. The variables included provide area information on the concentration of people born in the United Kingdom, the proportion of full-time workers, and the concentration of the population in a high socioeconomic class or with a high social grade. We also include information on LSOA population density and the concentration of individuals with high qualifications. Additionally, we include decile indicators showing the area's decile at a national level regarding income, education, health, crime, and living environment indicators, among others. Our preliminary results suggest that baseline male mortality is primarily influenced by area-level information related to the proportion of individuals in high socioeconomic classes, the proportion of full-time workers, and the concentration of individuals with high social grades, as captured by an indicator originally developed for market research and published in census data. For females, this component appears to be primarily influenced by the proportion of highly qualified individuals and full-time workers. The proportion of individuals who did not move in the LSOA in the last year and the proportion of individuals in the area who were born in the United Kingdom have minimal relevance. The results also suggest that the aging effect is influenced by multiple socioeconomic indicators with more subtle estimated effects. Information on deciles related to the living environment and crime seems particularly relevant for males, while information on deciles related to health is relevant for females, as well as other information, such as population density. In our work, we quantify these and other effects and exemplify how results can be interpretable even when machine learning models are used.

Keywords: mortality models - socioeconomic factors - longevity risk

References

[1] Cairns, Andrew J G, Torsten Kleinow, and Jie Wen. (2024). Drivers of Mortality: Risk Factors and Inequality. *Royal Statistical Society Series A: Statistics in Society.* **187**, pp. 989–1012.

[2] Wen, Jie, Andrew J G Cairns, and Torsten Kleinow. (2023). Modelling socio-economic mortality at neighbourhood level. *ASTIN Bulletin: The Journal of the IAA* **53**, pp. 285–310.

[3] Richards, Stephen J. (2020). A Hermite-spline model of post-retirement mortality. *Scandinavian Actuarial Journal* **2**, pp. 110–127.