Presentation by Adrian O'Hagan and Colm Ferrari

to the Society of Actuaries in Ireland on 15th January 2015

'Nonparametric and model-based clustering for actuarial data compression'

ABSTRACT

In the recent past, actuarial modelling has migrated from deterministic approaches towards the use of stochastic scenarios. Such projections are useful to an insurer who wishes to examine the distribution of emerging earnings across a range of future economic and mortality scenarios. The use of nested stochastic processes dramatically increases required computational time. This is particularly true for products with heavy optionality, which are becoming more popular in the marketplace. Incremental savings can be made as computing power expands and as coding is optimised. However, much more comprehensive savings are possible using a compressed version of the original data in the stochastic model. This involves the synthesis of "model points": a relatively small number of policies that efficiently represent the data at large. Traditionally this has been achieved using variations on the distance to nearest neighbour and k-means nonparametric clustering approaches.

Our research investigates how variations on this non-parametric approach, as well as an alternative model-based clustering approach, can be applied to actuarial data to produce high quality model points for stochastic projections. This is feasible since insurance policies typically have a number of associated location variables, allowing them to be modelled spatially. This is achieved using the standard Gaussian mixture model and automated using the freely available R packages Mclust and fnn (fast nearest neighbours). High quality historical data on a large set of 110,000 variable annuity policies has been provided by Milliman for the conduct of this research.

The nonparametric and model-based clustering approaches are contrasted with both the original weighted distance-to-nearest-neighbour approach and the outcome when the full, uncompressed data is used. The model points produced under each regime are compared for forecast accuracy at a range of compression levels for various stochastically generated economic scenarios. The results are validated using the Milliman actuarial pricing model.