Annals of Actuarial Science (2025), **19**, pp. 401–415
doi:10.1017/S1748499525100158

EDITORIAL

Research that I have not yet done, but that I consider worth doing

Alasdair David Wilkie

In QA Limited and University of York, York, UK
Email: david.wilkie@inqa.com

I start with two quotations.

On the scientific method:

"Furthermore, from the motion in latitude computed from our new method without those assumptions, we have proven that those very assumptions concerning sizes and distance are false, and have corrected them. We have done something similar with the hypotheses for Saturn and Mercury, changing some of our earlier, somewhat incorrect, assumptions because we later got more accurate observations. For those who approach this science in the true spirit of enquiry and love of truth ought to use any new methods they discover, which give more accurate results, to correct not merely the ancient theories, but their own too, if they need it. They should not think it disgraceful, when the goal they profess to pursue is so great and divine even if their theories are corrected and made more accurate by others beside themselves."

Ptolemy, *The Almagest*, IV, 9
(translation by G J Toomer, Princeton, 1998)

On random simulations:

"What might have been is an abstraction
Remaining a perpetual possibility
Only in a world of speculation."

T. S. Eliot, *Burnt Norton*, 1936

Having reached the age of 90, I realize (on the best actuarial grounds) that my future time to do the actuarial research that I would like to do is likely to be limited, so it is worth my while noting some ideas that I have had, and that others may like to take up.

These come under a number of headings, relating to the work that I have done over past years. These include mortality (and IP), financial indices, and financial modeling. I then add a historical note, which interests me because of the accidental coincidence of names.

Dr Şule Şahin has worked with me over several years on various topics, and after completing this paper, we plan to complete the research described in Section 1.1, then go back to the series of topics described in Section 3. How far we shall get with, these, if at all, we do not know.

The remaining research described in Section 1 is mostly for the CMI to consider or not as it wishes, though the items might also be interesting to an academic researcher if the data could be made available.

All the items in Section 2 are suggesting research or other work by FTSE-Russell, but I would still like to suggest them. Had I been still on the relevant actuarial or FTSE committee, I would also have suggested them there.

© Alasdair David Wilkie, 2025. Published by Cambridge University Press on behalf of Institute and Faculty of Actuaries. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

Section 4 is for anyone else with a historical interest to take up, if I have not done it myself.

If any researcher or student wishes to take up any of these ideas, and I am still available, I would be very pleased to explain further or assist.

Besides references to published works, I include some references to unpublished documents which I have placed on my website (https://davidwilkieworks.wordpress.com under Documents)

The quotation from Ptolemy above is relevant. He felt that others might be able to improve his models, and I expect that he would have been delighted to see the work of Copernicus, Kepler, Galileo, Newton, Einstein, and others. So, on a very much lesser scale, I am pleased to see any improvements in the modeling and forecasting of investment variables and of mortality.

I started programing in machine language for a Ferranti Pegasus computer in 1960. By 1970, I found that Fortran was more useful for research work, and I have used it in different versions since then. I found various mathematical subroutines in an IBM manual (IBM, 1970) and in various other sources, including Press et al. (1992), NAGLIB, Collected Algorithms of the ACM, etc. These I have copied and sometimes amended. I have thus developed a suite of Fortran subroutines for many purposes, and I could make these available to anyone who was interested (with no guarantees about accuracy, though I believe them to be correct). Newer versions of any of these publications may give more algorithms.

For some years, I have found it convenient to start a program in Excel using VBA to read in parameters and controls, passing these to a Fortran DLL, which does all the calculations, and passes the main results back to VBA for display, perhaps with graphs in Excel. I may use Fortran to read in large data files and output in "csv" format all the intermediate results necessary to check the calculations. Others will choose different methods.

Most of these research ideas require data. I would expect that the CMI would co-operate very readily with any genuine researcher for mortality and IP data. Much of the investment data, however, comes from a variety of published sources that I have followed myself for many years. I describe these in Documents: *Data Sources*, in which I give further references to data files that I have constructed for myself.

Note: Sections with two digits, e.g. 1.1, describe research topics. Sections with three digits, e.g. 1.1.1, give additional comments on the Section above.

IBM (1970) Ststem/360 Scientific Subroutine Package Verion III Programmer's Manual

Press W H, Teukolsky S A, Vetterling W T., and Flannery B P. *Numerical Recipes in C*. Cambridge 1992. (There is a version in FORTRAN, but C is easy to translate into Fortran and vice versa).

1. Mortality

1.1 Complete our current Covid investigation

The next task for Dr Şule Şahin and me after I have completed this paper is to complete our investigations into forecasting Covid mortality, allowing, or not allowing, for the "Spanish" flu of 1918. Our preliminary results have been presented in the AFIR-ERM Pandemic Working Party Report on Mortality, but it was clear at that time that the distributions we were using were not fat-tailed enough, and that we needed to investigate others.

Our method was to choose countries with satisfactory data, not affected by the two World Wars, with one starting date before 1918 and another after that date, ending in 2019, using the Lee-Carter method first, then fitting a time-series model to the time parameter, $\kappa(t)$, with some fat-tailed distributions (described below). Then we used a hypermodel (also described below) to forecast mortality in 2020, which we could then compare with the actual experience for 2020 (including the Covid experience).

Once we have tried other distributions that give better results for this first stage, we then expect to add the actual data for 2020 and a few subsequent years, to give stochastic models allowing for both pandemics.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

1.1.1 Distributions

The distributions we used in the first stage of our Covid investigations (and have also used in some investigations of our investment data, not yet published) were what we call the "conical" series: Normal, Laplace, Skew Laplace, Hyperbolic, and Skew Hyperbolic (see Documents: *Notes on Conical Distributions*). However, the skewness and kurtosis of the conical distributions are limited, and some of our data (both mortality and investment data) show much higher kurtosis than the conical distributions allow (see Documents: *Notes on Other fat-tailed Distributions*).

Following Fernandez and Steel (1988), we add a Skew Normal (not immediately useful here on its own). We also add a Double Pareto, Skew Double Pareto, Student-t, and various mixes of some of these.

Our mixes are in the following basis: assume one random variable X_1 with some given distribution, and another X_2 also with some given distribution, not necessarily the same. Then another random variable Y equals X_1 with probability p, and equals X_2 with probability $q = 1-p$. We can use Y for our residuals and innovations. We prefer to use the same location parameter, μ, for X_1 and X_2 because none of our data shows any bimodality.

Calculating the likelihood function of Y and its moments is straightforward, as is simulating random values of Y. The kurtosis of Y can be very large.

But (looking ahead to our hypermodel), we note that the likelihood function of the Laplace distribution (symmetrical or skew) is not differentiable with respect to μ, and we have to use finite difference approximations. The same is true of some other distributions, but a mix of normals is fully differentiable, which gives a minor reason to prefer this if it is otherwise suitable.

Fernandez and Steel (1988) on Bayesian Modelling of Fat Tails and Skewness, *Journal of the American Statistical Association*, Vol 93, No 441, pp 259–371.

1.1.2 "Hypermodels"

I often use my models for random simulation of the future, for example, for annual values of investment or mortality variables. The basic simulation will allow perhaps for a time-series model with fixed parameters, and with random values of the annual innovations for each variable according to some distribution with fixed parameters.

However, if one estimates the values of the parameters of any model from limited data, there is always some uncertainty about the estimated values, and I allow for this parameter uncertainty through what I call a "hypermodel." I treat the parameters also as set of random variables, each distributed normally, with a given vector of means and a covariance matrix, better expressed as a vector of standard deviations and a correlation matrix. I convert the correlation matrix to its Cholesky equivalent. Then, before each simulation for a number of years, I can generate values of the parameters for use in that simulation, using a vector of independent normally distributed variates.

I generally use maximum likelihood estimation to obtain estimates of the parameters, which can be treated either as fixed (in the basic model) or as the parameter means (in a hypermodel). The covariance matrix of the parameters can be obtained by inverting the negative of the matrix of second partial derivatives of the likelihood with respect to each of the parameters (noting also that at the maximum, the vector of first partial derivatives should be zero).

Since in the hypermodel we assume that each parameter value is normally distributed, it is desirable to use transforms for certain of the parameters so that the resulting transform may have a full range. The three transforms I have used are (denoting the parameter as p and the transform as t):

(1) If $p > 0$, then $t(p) = \ln(p)$ and $p(t) = \exp(t)$. This is needed for scale parameters within distributions for innovations, such as σ for Normal and λ for Laplace or Pareto.

(2) If $0 < p < 1$, then $t(p) = \ln(p/(1-p))$ and $p(t) = \exp(t)/(1-\exp(t))$. This is needed for some time series parameters, where we wish an autoregressive parameter to be limited to give long-term stability.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

(3) If $-1 < p < 1$, then $t(p) = \ln((1+p)/(1-p))$ and $p(t) = (\exp(t)-1)/(\exp(t) + 1)$. This is needed for some skewness parameters and possibly other time-series parameters.

Transforming the parameters does not change the location of the maximum in terms of the new parameters, so this is not a problem.

1.2 A financial problem if annuity/pension mortality depends on amount

Some years ago, the CMI Self-Administered Pension Schemes investigation noticed that the mortality of those drawing pensions varied by the amount of the pension, being lighter when the amount was higher, and heavier when it was smaller. They split the data into three bands: small, medium, and large, which had heavier, medium, and lighter mortality.

It was easy then to show that if these bands were used for pricing, then there were anomalies. If someone applied with a very small purchase price, they clearly got a small pension and the high mortality was used for them. But as the purchase price increased, the pension purchased increased, and at some critical point, the amount of pension fell into the medium band; so one then should use the medium level of mortality, which gave the applicant a smaller pension, which fell into the small band. At a higher level of purchase price, the amount remained in the medium band, but there was an ambiguous layer in between. The same problem arose as one moved from medium to high. It was all very well using the differential mortality for reserving but it would not do for pricing.

Another method, which was tried out by a student at Southampton University for a piece of work for a master's degree, under my supervision, was to fit a mortality rate that varied continuously by amount. We received the data from the CMI, summarized in very narrow amount bands, not as individual records. It was clear that above some large amount, the mortality rates were not distinguishable from those for moderately high amounts, and also that those who got very small pensions had mortality much the same as those with moderately small ones. So it was worth trimming the amounts to put those below some lower level and those above some upper level at fixed lower and upper points, which in this case were £3,000 and £35,000; but these numbers would need investigating in any particular case.

Thus A was defined as: $A = \text{Min}(\text{UL}, \text{Max}(\text{LL}, \text{Amount}))$

For the base mortality, we then chose a simple Gompertz model, using a transform t of x (age): $t = (x-72.5)/22.5$ so that t ranged from -1 to $+1$ as age ranged from 50.5 to 95.5. This formula could have been extended with higher terms in Chebyshev polynomials as used in my GM(r, s) models. But the "central" mortality rate for some central amount (£5,000 in this case) was then:

$$\mu(x) = \exp(\beta_0 + \beta_1.t).$$

We then saw that the variation in mortality from low to high was proportionately much bigger at lower ages than at higher ages, so we need to reduce the amount of variation as age increases. So the amount was then brought in by multiplying this central rate by:

$$1 + \beta_2.f(A).\exp(-\beta_3.t)$$

where $f(A) = (A - 5{,}000)/10^7$

This gave us four parameters. We estimated their values by the maximum likelihood method with a Poisson assumption for the distribution of deaths in each age/amount cell. The estimated values of the parameters were: $\beta_0 = -3.621539$, $\beta_1 = 2.482734$, $\beta_2 = 58.422614$, and $\beta_3 = -0.009228$.

All this fitted the mortality reasonably well. However, we did not go the further step to calculate annuity values for each narrow amount band, to see whether the function was monotonic in an appropriate way, or whether there still were anomalies. This research would be worth doing. While

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

the model and parameters given above could be used for this, it would also be desirable to fit this model, or a similar one, to more up-to-date data.

1.3 The negative binomial (Polya) distribution

It has often been found that in a mortality investigation, the deviations of actual from expected deaths are not distributed as Poisson but are fatter-tailed. This is sometimes dealt with by assuming an "over-dispersed" Poisson. However, this is not a real discrete distribution but is a normal distribution with the variance not equal to the mean (as it would be for the normal approximation to a Poisson) but rather larger by some over-dispersion factor.

An alternative would be to assume a negative binomial distribution (see negative binomial distribution – Wikipedia) for the numbers of deaths (or indeed of other events such as falling sick or recovering from sickness or incurring a critical illness event). The negative binomial distribution is also called a Pascal or a Polya distribution. For brevity, I shall call it a Polya distribution. It has two parameters p and r, and its probability function is:

$$p(k) = {}_{k+r-1}C_k.(1-p)^k.p^r$$

If r is integral, it is the probability of the number of failures (k) before r successes in a Bernoulli test. But r need not be integral. The distribution can also be derived by assuming that in a population each individual has their own Poisson parameter, which is distributed according to a Gamma distribution. Although the Gamma gives an exact result, I imagine that a Lognormal distribution for the Poisson parameter might give a similar result. This could represent a population with differing degrees of frailty or sturdiness.

The mean is $r(1-p)/p$ and the variance is $r(1-p)/p^2$.

However, with most mortality investigations, we get only one sample at each age with the number of years exposed and the number of deaths, and this does not give us enough to estimate two parameters for each age. There is another way of investigating the data that I suggested to the CMI some years ago but never completed.

This method was to take a file with individual records, which gave us, for each individual, exact dates of birth and of entry and exit from the investigation, so we could go through the file and, for each individual, allocate the exact number of days within each age that the individual passed through, and whether or not the period terminated in death. Then the innovation would be to collect data for each age in boxes of say 100 years or perhaps 35,000 days, recording the time exposed and the number of deaths within each box. We could then look at the distribution of deaths within all the boxes within one age to see whether a Poisson or a Polya assumption gave the better fit, and we could estimate the values of p and r for each age.

We would then like to smooth (graduate) in some way the values for each age, which could now be called p_x and r_x.

At younger and older ages, there would probably not be enough data to give many boxes within each age, and we might need different sized boxes at different ages depending on the numbers of deaths.

It is essential that the entries in the data file are randomly arranged to start with. In the file I looked at in the past, the entries were arranged so that, within each pension scheme, all the deaths were shown, then all the survivors. This would not do. I would then have generated random integers within the range -2^{31} to $+2^{31}-1$, attached one integer to each record, then sorted the records in numerical order by the attached integer. This would be straightforward in Fortran but I don't know how easy this would be in R or some other system. (For more about generating random integers, see Documents: *Some aspects of random number generation.*)

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

1.3.1 Sorting methods

When doing a large number of simulations, one wishes to obtain certain statistics from the results. It is easy to accumulate moments as one goes through, but if one wishes to get quantiles, one needs to record all the values and sort them into numerical order at the end. In one exercise, I had done 10,000 simulations with no trouble, but when I moved up to 100,000, I found that the sorting process was taking very much longer than generating the simulated values.

I then looked up Knuth (1973) and found that my method (a "Duby" sort) was $O(n^2)$, whereas Quicksort (See Quicksort – Wikipedia) was generally $O(n.\ln(n))$. I then tried generating one million random numbers and sorting them. Using my old method took 100 seconds, and using Quicksort took 600 milliseconds. I should have had a much better improvement, since $1{,}000{,}000/\ln(1{,}000{,}000) = 62{,}040$, but there would be an overhead in noting and writing out the time.

This showed that the sorting algorithm you choose is not a trivial matter. My present practice is that if the number to be sorted is very small, say no more than a few hundred, any method will do. If the number is in tens of thousands, Quicksort is best. However, if the number of simulations is too big to fit into the computer store at one time, one needs to use merges.

The CMI file in the case had (for males only) over 2,000,000 records. I would have sorted the first 300,000 and written them to an output file A, then the next 300,000 to file B, the next 300,000 to file A, and so on alternately. In the second pass, I would take the first files in A and B, which are sorted in order, and merge them to file C, then the next pair to file D, and so on, reversing between A/B and C/D, until one had only two files, which could be merged to become one.

I also now use two programs, one to generate the simulations, the other to sort them and extract the statistics, quantiles, etc, as desired.

There is much more in the computer literature about sorting.

Knuth D *The Art of Computer Programming Vol 3, Sorting and Searching*. Addison-Wesley 1973.

1.3.2 Data compression

If the file one wishes to process is very large, there may be advantages in compressing the data into a much smaller space than it may be originally provided. The CMI file in question had over 2,000,000 records in "csv" format, with 53 characters (or 53 bytes) per record. This was rather large.

We can note that sex is only Male or Female (in this file), so only one bit is necessary to record it. Whether the pensioner dies or not is similarly only one bit, and does not need a long word. Dates were given in full with hyphens e.g. 2004-01-01 or 10 bytes each. The first and last dates in the investigation were 1/1/2004 and 31/2/2009; these are Excel dates from 37,987 to 40,178, so by subtracting 37,987 from each date, the maximum date was 2,191, so any such date could be expressed in four decimal digits or no more than 12 bits. Dates of birth had a much wider range, but an Excel date in this or the last century is no longer than 5 digits or 16 bits.

The commas in a csv file help one to read it, but do not give any information when items can be put in a fixed format, so they can be omitted.

Once I had looked at the number of bits necessary to express each item, I could combine them into a long integer by multiplying by powers of 10, powers of 2, or a mixture. I found that I could get the original 53 bytes into less than two 32-bit integers or less than 8 bytes. This would have made the sorting very much quicker.

A compressed file may be difficult for a person to read, but with correct compression and decompression subroutines, they are straightforward to process.

It is only rarely that I have found this sort of data compression necessary, but many insurance files may contain very large numbers of cases, and actuaries may like to think about how they can be most efficiently processed.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

1.4 IP analysis and valuation by cause of sickness

In a number of papers, Sing Ye Ling et al. (2010) took the Income Protection (IP) data investigated by the CMI and subdivided it by cause of sickness, grouping causes with few claims. She fitted graduated curves to the recovery rates and mortality rates of claims, subdivided by cause. She also found a distribution of the numbers of claims occurring by each cause or cause group.

It would then have been a straightforward calculation, which I had no time to do, to carry out a valuation of the claims in force in the CMI IP file at a year-end, allowing for some pattern of claim payment (e.g. end of each month plus a fraction for odd days), both on an aggregate basis, not subdivided by cause, and on a basis allowing for each claim separately.

This could also be done, confidentially, for each contributing office separately. This would give an overall view as to what the benefit would be of valuing outstanding claims in aggregate or separately by cause of claim.

I remember, when the new methods to replace the "Manchester Unity" method were being introduced by the IP (then PHI) Committee of the CMI in about 1980, one actuary from a PHI office commented that he did not like the new methods. If he knew that there was a big influenza epidemic at the year-end, he would know that many claims would recover quickly, so treating all claims the same way was not appropriate. The investigation I suggest would have suited him perfectly.

S. Y. Ling. H. R. Waters and A. D. Wilkie. (2010a) "Modelling income protection claim terminations by cause of sickness, I: recoveries" *A.A.S.*, **4**, 199–239.

S. Y. Ling. H. R. Waters and A. D. Wilkie. (2010b) "Modelling income protection claim terminations by cause of sickness, II mortality of UK assured lives." *A.A.S.*, **4**, 241–259.

S. Y. Ling. H. R. Waters and A. D. Wilkie. (2010c) "Modelling income protection claim terminations by cause of sickness, III: excess mortality" *A.A.S.*, **4**, 261–286.

2. Investment indices

2.1 Bias in the FTSE 250 Index

It has been noted that the FTSE250 index has performed much better than the FTSE100 index over a number of years. The best data that I have available when writing this is for the 22-year period from end-June 2002 to end-June 2024. Over this period, the FTSE100 increased from 4,656.36 to 8,164.12, a ratio of 1.7533 or about 2.59% per year, whereas the FTSE250 increased from 5,496.64 to 20,286, a ratio of 3,6806 or about 6.12% per year. This shows a considerable difference.

Both indices contain stocks registered in the UK, but the FTSE100 index includes a lot of international stocks, doing business around the world, whereas the FTSE250 includes mainly stocks doing business in the UK. It is suggested that the variation in index performance shows that the latter (smaller) stocks do better than the former (larger) stocks.

I suggest that there is also a bias in the construction of the two indices, which may account for part of the differential. The stocks in the FTSE100 index are not the largest (by market value) on each day, and those in the FTSE250 index are not the next 250 on each day. There are rules for the process, which are laid down by FTSE-Russell. The normal procedure is that on a particular day in each quarter, any stock in the FTSE250 that is in a position above 90th (by market value) is promoted to the FTSE100, and any stock in the FTSE100 that is below the 110th is demoted, along with enough others nearby to keep the total number of stocks 100 and 250.

A similar process is carried out between the FTSE250 and "the rest."

There must also be exceptions when new stocks of large enough size appear and disappear by new issues, spin-offs, insolvencies, etc, but I do not look at these.

If we look at the normal rules and compare them with hypothetical "exact" indices that change every day, we see that, when a smaller stock gets bigger, its rise from position 110 to 101 "correctly" goes into the FTSE250 index, but its rise from 100 to 90 also goes into the FTSE250 index,

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

"incorrectly." When a large stock gets smaller, the same happens in reverse, its drop from 101 to 110 goes into the FTSE100 index, also "incorrectly."

Since it is the largest stocks in the FTSE250 that behave "incorrectly," they may make a big difference. In the FTSE100 index, it is the smallest stocks, so it may make less difference. The same is true between the FTSE250 and "the rest". I do not have daily data of the prices and other details of stocks, and I don't know who would have, other than FTSE-Russell itself. But it would be a useful investigation to run parallel indices, the "exact" ones, for say a year, and see what the bias is.

2.2 The Gini coefficient for the distribution of sizes of stocks

P. E. Hart (1959) wrote a paper about the sizes of companies in the UK. What prompted the paper was the observation that there was a belief that large companies were getting larger, meaning that there was an increase in business concentration. At that time, the two largest companies in the UK were Unilever and ICI; the former still exists, and the latter also in the form of AstraZeneca, its descendant, and both are still large.

The author looked at the concentration of companies in the past. He chose dates 1908–10, 1924, 1938, and 1950. He used company profits rather than market value as a measure, and he included non-quoted stocks as well. He limited it to manufacturing companies, so excluded financials, oils, and mining. He found that although there were changes in the overall distribution of size from time to time, there was no constant move toward higher concentration.

The particular companies had changed considerably, though some were durable; some have changed names, which those without historical knowledge may find confusing. What I found interesting was that the largest company in 1908–10 was J&P Coats of Paisley, which produced Anchor thread. It had subsidiaries around the world, particularly in Russia, but these had disappeared by 1924.

The usual statistical comparator of the distribution of sizes of many statistics, such as income, population size, etc, is through the Gini coefficient (see Gini coefficient — Wikipedia). It would be quite possible for FTSE-Russell, in the calculation of all the major FTSE indices, All-share, FTSE100, and FTSE250, FTSE350 etc (but not the individual sector indices which perhaps have too few stocks) to calculate the Gini coefficient and publish it each day. After it had been explained to people, it would interest them.

What would also be of interest would be a "league table" of the companies in the FTSE100 (or even the FTSE350) index, showing their relative positions by market value, say at the end of each quarter.

A more complicated job for an economic historian would be a history of the large companies in the UK (and equally in other countries), showing the rise and fall of what from time to time seem great names. Brunner Mond was large in 1908–10, became (along with the British Nobel company and others) Imperial Chemical Industries (ICI) by 1924, and part of it is now left in AstraZeneca. Guinness was large in 1908–10 and 1924, Distillers was large in 1938 and 1950, and the combination, Diageo, is large now. Hanson has come and gone. Others, like the big textile and motor companies, have gone.

Hart P.E. (1959) "Business Concentration in the United Kingdom", *Journal of the Royal Statistical Society, Series A*, Vol. 123, No. 1 (1960), pp. 50–58

2.2.1 A practical example of Gini coefficients

The *Sunday Times* newspaper publishes every Sunday a list of the 200 largest companies, by market value, on the London Stock Exchange, so it is quite possible to calculate the Gini coefficient for all or subsets of these. I use the data given on 18 August 2024.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

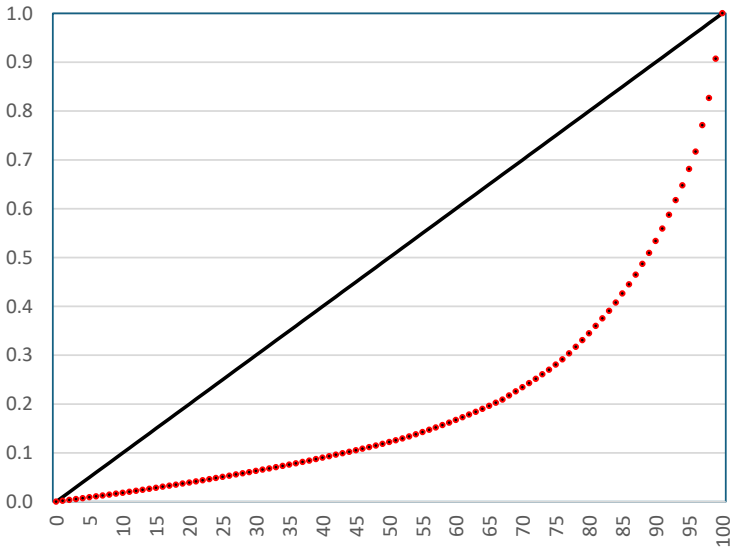


Figure 1. Gini coefficient.

Figure 1, shows the principle, using the top 100. The horizontal axis represents the 100 stocks from smallest to largest. The vertical axis shows the successive sums of the market values as a proportion of the sum of all companies.

In symbols, let there be n stocks. Let the market value of stock i be V_i, with each $V_i \leq V_{i+1}$, so that the stocks are in increasing order of size. Let the sum of the market values of stocks 1 to k be $S_k = \Sigma_{i=1,k}\ V_i$, so that the sum for all stocks is S_n. Then the horizontal axis shows i and the vertical axis shows $P_i = S_i/S_n$.

In Figure 1 the graph of S_i is shown in red dots. The area to the lower right of that, up to the axes, has area A. The area above the red dots and below the diagonal line has area B. The total area below the diagonal is area $C = A + B$, and the Gini coefficient, g, is defined as B/C.

We can calculate the area A by filling in between the dots linearly. Then A can be calculated as the sum of horizontal stripes:

$$A = \Sigma_{i=1,n}(n - i + 1/2),\ V_i/S_n$$

The lowest horizontal stripe represents the contributions from stock 1, with 99 little rectangles across the bottom, plus a triangular half at the left-hand end. The highest stripe is just $\frac{1}{2}V_n$ at the top for the largest stock n.

The area $C = n/2$. Then the Gini coefficient $\mathrm{g} = B/C = (\frac{1}{2} - A)/C$.

For the largest 100 stocks in the *Sunday Times* data, the range is from £3.8 billion to £202.2 billion; the total market value is £2,169.7 billion; and the Gini coefficient is 0.594. The results for other sets are shown in Table 1.

Observe that the Gini coefficient gets larger when one adds more stocks at the bottom or at the top. The values are not significant in themselves, but they would be relevant if they were shown for every day, and one can see the drift over time within one group. One could also compare indices for sectors or for different countries if the numbers of stocks were the same.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

Table 1. Results for Gini coefficient calculations

Group	Smallest £ b	Largest £ b	Total £ b	Gini Coeff
Top 50	7.7	202.2	1,905.3	0.475
Top 100	3.8	202.2	2,169.7	0.594
Top 150	1.8	202.2	2,302.6	0.651
All 200	1.2	202.2	2,375.6	0.693
Bottom 50	1.2	1.8	73.0	0.074
Bottom 100	1.2	3.7	205.9	0.195
Bottom 150	1.2	7.7	470.3	0.308

2.3 The annual reports on the indices

For many years, annual reports on the FT-Actuaries (later FTSE-Actuaries) share indices appeared in *Journal of the Institute of Actuaries* and *Transactions of the Faculty of Actuaries* (*TFA*). They were written at first by Jack Plymen and later by others, the last being John Brumwell, up to 2002 (Brumwell, 2003). When the FTSE-Actuaries gilt indices started, I wrote similar annual reports for them for a number of years up to 1984 (Wilkie, 1984–86), but my successors who had the responsibility of scrutinizing these did not continue them. Now the annual reports on the share indices have also stopped.

These reports were informative and useful from a historical point of view. It would be useful for FTSE-Russell to continue them and to publish them more widely than in actuarial journals.

A full record of publications on the indices, prepared by David Raymont, is in Documents: *Papers on the Actuaries and FTSE Actuaries Indices*.

Brumwell, J.C.H. (2003). *Notes on the FTSE Actuaries Share Indices (United Kingdom series) in 2002. BAJ (2003)* ***9(2)****: 457–480.*

Wilkie, A.D. (1984—1986). *Notes on the Financial Times-Actuaries Fixed Interest Indices up to 1984. TFA (1984—1986)* ***39****: 587–626.*

3. Investment

The main line of my investigations is in improving the "Wilkie model." I shall observe here that many commercial providers of "Economic Scenario Generators" seem to have used some version of my model, or been inspired by my model, but have not published their improvements, and have kept them confidential. This is not a good way of making scientific advances, and it would not have pleased Ptolemy.

3.1 Complete investigation by fat-tailed distribution and hypermodel

Before the COVID-19 pandemic interrupted us, Dr. Şahin and I were working on elaborating the Wilkie model in two ways: using fatter-tailed distributions for the innovations, and constructing a hypermodel. We have done this in the mortality modeling noted in Section 1.1.

Our investment model, however, has added complications. To get the covariances in a hypermodel, we need to estimate all the interconnected variables together simultaneously. If one considers the model for retail price inflation, the residuals $QE(t)$ do not appear in the model for any other variables, so that variable can be estimated independently. The actual values are used

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

elsewhere, but the model has no effect. In other cases, such as share yield, $Y(t)$, share dividend, $D(t)$, and long-term interest rate, $C(t)$, there are dependencies on $YE(t)$ and the total likelihood must be estimated jointly.

When we introduce different distributions for the innovations, the time series parameters estimated by maximum likelihood may also change, sometimes very little, but sometimes by quite a lot. It would be very laborious to try all the possible combinations of distributions for each of the possible variables, and one has to derive a practical method of getting suitable answers, like finding the most suitable distribution for the innovations for each variable separately and then combining them without looking for other distributions.

A further problem is that the time spans of our variables differ. The ones used in my earliest, 1984, model have data from 1923; others start only in 1962, when earnings on the share indices become available; a further variable, the yield on index-linked bonds, starts only in 1981. If one wishes a hypermodel for all the variables combined, one may have to estimate the covariances for the longer series first and then slot in the additional covariances for the shorter series, including only for the new variables. This may or may not produce a valid covariance matrix, and this has to be checked.

In aggregate, there is lot more work to do with the investment series, but it is next on our agenda after we have finished the mortality modeling.

3.2 Update model to June 2025 and compare with year-ends in different months

In Section 3.1, we would of course use the latest data up to June 2025 or whatever the latest year is. When I started modeling in the 1980s, I chose June for the year-end because it avoided the extreme value of the dividend yield at the end of 1974. This was done to make it more likely that a normal distribution would be appropriate. However, if we are introducing father-tailed distributions, it would be appropriate to try each separate month as the year-end for the modeling and see what difference it makes to the time series parameters and to the distributions of innovations.

3.3 How to introduce fatter-tailed distributions into our stochastic bridging model.

In the model for stochastic bridging, which we described in Wilkie and Şahin (2016a, 2016b, and 2016c), it was clear that the monthly residuals in our stochastic bridging model were often fatter-tailed than normal. But if we take a fat-tailed distribution, say Laplace, and add a number of Laplace-distributed innovations together, the sum is not Laplace-distributed. Unless the variance of the distribution is infinite, the sum of the innovations will tend towards normality. And it may be that the annual innovations are fat-tailed. I am not sure how to deal with stochastic bridging in these circumstances, but perhaps some other researcher has faced this problem. It needs further investigation.

Wilkie, A.D. and Şahin, Ş. (2016) Yet more on a stochastic economic model: Part 3A: Stochastic interpolation: Brownian and Ornstein–Uhlenbeck (OU) bridges. *Annals of Actuarial Science*, Vol 11, Part 1, pp 74–99.

Wilkie, A.D. and Şahin, Ş. (2016) Yet more on a stochastic economic model: Part 3B: Stochastic bridging for retail prices and wages. *Annals of Actuarial Science*, Vol 11, Part 1, pp 100–127.

Wilkie, A.D. and Şahin, Ş. (2016) Yet more on a stochastic economic model: Part 3C: Stochastic bridging for share yields and dividends and

interest rates. *Annals of Actuarial Science*, Vol 11, Part 1, pp 128–163.

3.4 A new model for exchange rates

At the AFIR Colloquium in Madrid in 2011, I presented a new model for exchange rates between currencies (see Documents: *AFIR 2011 Madrid Wilkie*). My earlier model for exchange rates

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

(Wilkie, 1995) was based on purchasing power parity. If $XR_{ij}(t)$ is the number of units of currency j for one unit of currency i, and $Q_i(t)$, $Q_j(t)$ are the price indices in i and j, then I defined:

$$X_{ij}(t) = \text{In}(XR_{ij}(t)) + \text{In}(Q_i(t)) - \text{In}(Q_j(t))$$

and gave $X_{ij}(t)$ a standard AR(1) model:

$$X_{ij}(t) = XMU_j + XA_j \times \{X_{ij}(t-1) - XMU_j\} + XE_j(t)$$

This fits adequately for one country alone, but with several countries, the cross-rates between j and k become very messy.

In my new model, I assume a hypothetical or "hidden" series for each country, $HR_i(t)$, representing the "relative strength" of the currency.

Then I put:

$$S_i(t) = Q_i(t)/HR_i(t)$$

and set:

$$XR_{ij}(t) = S_j(t)/S_i(t) = Q_j(t)/Q_i(t).HR_i(t)/HR_j(i)$$

Then:

$$\ln(XR_{ij}(t)) = \ln(Q_j(t)) - \ln(Q_i(t)) + \ln(HR_i(t)) - \ln(HR_j(t))$$

I put:

$$H_i(t) = \ln(HR_i(t))$$

and fit a time-series model to $H_i(t)$.

But I do not know the values of the hidden series $H_i(t)$. So, I took 12 currencies and a long series of dates, and on each date fitted the values by least squares. I had to use the mean values of each $\ln(Q_i(t))$ and each $\ln(XR_{ij}(t))$ over the relevant time period.

I found that if one omitted one country entirely, the relative values of the hidden series did not change, but if one estimated over a different time period, the means were changed, and thus all the values of the hidden series.

I then observed that if the monthly series were AR(1), then the annual series would also be AR(1), and I estimated some values. I noted high cross-correlations between the residuals for different countries, and also very high kurtosis for certain countries. However, I have never completed and written up in a paper on my investigations, which I would now like to update for the added years.

Wilkie A.D., 1995, More on a Stochastic Aset Model for Actuarial use, *BAJ*, Vol 1, pp 777–964,

3.5 Analysis of the yield curve, with new parameters and time series analysis of them

Until 1995, interest on UK government stocks (gilts) was taxed, but capital gains were not. This resulted in there being a considerable variation in gross redemption yield depending on the coupon rate, and in the FTSE-Actuaries indices, there were yield curves for low-, medium-, and high-coupon stocks. In 1995, the tax system changed, and a single yield curve appeared.

In 2012, FTSE invited Andrew Cairns and me to produce a new formula for the gilts yield curve. We did so, using data from 1995, and described the results in Documents, *The New Yield Curve*. The formula chosen for the forward rate for term t on any day was:

$$f(t) = b_0 + b_i.\exp(-c.t) + b_2.\exp(-3c.t) + b_3.exp(-5c.t) + b_3.\exp(-7c.t)$$

The value of the parameter c was fixed at 0.04, which seemed to give good fit over the available period. The values of b_0 to b_4 vary from day to day. From the forward rates, values of the zero-coupon rates can be calculated and from these "par yields" which are published daily in *The Financial Times* for selected terms.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

I had hoped to investigate further the fitting of this formula, using perhaps a sliding value of c which could vary slowly from day to day, but was not allowed to jump around, as might be possible if the value was fitted separately each day; but this has not been done.

My next thought, which could usefully still be done, was to treat the daily values of b_0 to b_4 as time series, which would give a yield curve for the Wilkie model, and perhaps replace the values of $B(t)$ and $C(t)$, my short-term and long-term interest rates. The value of b_0 is the forward rate for infinite t, but is not the par yield for a perpetuity (my original C), and the sum of the five bs is the forward rate for term zero, but this is not the same as the current bank rate (which I use for B).

The daily values of the five b parameters are not published by FTSE-Russell, but it should be possible to derive them from the five published yields, though I have not tried this. The client files give par yields, zero-coupon rates and forward rates for 10 durations, and it should be possible to get the values of the parameters from these with some accuracy.

In the current Wilkie model, the model for $C(t)$ depends on retail price inflation $I(t)$, and there is a connection with the residual for the Share Dividend Yield, $Y(t)$. One would need to investigate these too.

3.6 Reconsider model for long-term interest rates, C, with inflation

In the present Wilkie model, the value of $CR(t)$ is logged so that the "real" yield in excess of smoothed inflation must be positive. However, in recent years, the value of $C(t)$ has often been less than the value of inflation, however that is smoothed, so it looks as if CR can be negative at times. We found the same problem with the index-linked yield, R, which in my 1995 model was assumed to be always positive, so $\ln(R)$ was modeled. But the yields on index-linked stocks have subsequently become negative, so it was necessary to change the model for the index-linked yield to model R directly.

It would therefore not be unreasonable to change the model for C to allow negative real yields. This should be fairly straightforward to investigate and it is consistent with the yield curve described in Section 3.5, where the formula allows negative yields.

Şahin et al. (2014a and b) used different data (from the Bank of England) to do something very similar, but I suggest here using a different source.

Şahin, Ş., Cairns, A., Kleinow, T. and Wilkie, A. D. (2014a) A Yield-Macro Model for Actuarial Use in the United Kingdom, Annals of Actuarial Science, 8(2), 320–350.

Şahin, Ş., Cairns, A., Kleinow, T. and Wilkie, A. D. (2014b) A Yield-Only Model for the Term Structure of Interest Rates, Annals of Actuarial Science, 8(1), 99–130.

3.7 Try fitting dividends using XD adjustment figures

In the late 1970s, the government controlled share dividends, which were, for a period, not allowed to rise, but had to be held as retained earnings. In the 1980s, a new government removed these restrictions, so many companies issued special dividends to allow for the backup. In the FTSE-Actuaries indices, only "normal" current dividends are included, but not "specials." I therefore suggest to the then Index Committee that the actual dividends should also be recorded. This in due course resulted in the "xd adjustment," which gives the quantity of dividend (or interest for gilts) going "ex dividend" on that day, and it is the xd adjustment that is allowed for in total return indices.

The xd adjustment for the share indices is now available from 1985. It would therefore be possible to use a 12-month moving total for this in place of the dividend index, $D(t)$, that I have used in the Wilkie model, which up to now has been based on the share price index, $P(t)$, multiplied by the dividend yield, $Y(t)$.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

There is also a difference in timing between these two versions of the index. The published yield takes into account changes very soon after they are announced, but the xd adjustment changes on the ex-dividend date, which is always some time later.

3.8 Refit overall model to other countries

In my 1995 paper, I fitted my then model to several other countries. The model has been updated since then but only for the UK, and it would be nice to see how well it fits the data for other countries. Since I am suggesting other improvements in these notes, it would be desirable to fit the best available UK model to other countries as well. Obtaining data may be a problem, but the OECD website has data for price inflation and interest rates, and share price data (and derived exchange rates) can be obtained from the FTSE World Indices. There may be other possible sources that I do not know about.

4. Historical

4.1 The reverend David Wilkie

The Reverend David Wilkie was born in 1738 and died in 1812. He became minister of Cults, a parish in Fife. Among his children was another David, who became Sir David Wilkie RA, one of the most famous artists in Britain in his day, though nowadays a little neglected. The minister also took an interest in actuarial matters.

He prepared a life table for the neighboring parish of Kettle in Fife, which was published in the Statistical Account of Scotland in the 1790s (see Documents: "ParishOfKettle-SAS"). He wrote a book, *The Theory of Interest*, which was published in 1794, and was said to be the bedside reading of the then Prime Minister and Chancellor, William Pitt (but I do not know where I read this). And there is a record of his having been consulted in 1805 by the instigators of the Friendly Society of Teachers in Fife (1812).

There is a portrait of the minister and his wife, done by his son, in the National Gallery of Scotland. There is a memorial in Cults Parish Church to the minister and his wife done by Sir Francis Chantrey, the leading British sculptor of his day, which was commissioned by the artist (who would have known him as a fellow RA). It would be nice to include good photographs of both of these in a fuller paper.

The only source I have seen for his life is in Alan Cunningham's book on the artist (1843). It is of course a little confusing to have so many David Wilkies around and I would distinguish them as the minister, the artist, and the current actuary. I have long meant to write a note on the minister's life and work and review his very comprehensive book, but this has not been done yet. Perhaps someone else would like to take it up.

Although I have the same name as the minister and his son, I do not appear to be related. My great-grandfather was born in Brechin in Angus about 1820. The minister was born in Ratho Byres, then in Midlothian, now on the border between the City of Edinburgh and West Lothian, where his ancestors were said to have farmed for four hundred years. There is a village of Wilkieston nearby and a Wilkie's basin on the Union Canal also nearby. But all this is a long way from Brechin, and there are many Wilkies there too.

Cunningham, Allan (1843), *The Life of Sir David Wilkie, 3 Volumes*. London.

Parish of Kettle in *Statistical Account of Scotland*, 21 Volumes, 1791 to 1799.

Wilkie, the Reverend Mr David (1794), *Theory of Interest, Simple and Compound, derived from First Principles, and Applied to Annuities of all Descriptions*. Edinburgh.

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press

Friendly Society of Established Teachers in Fife and other Counties in Scotland (1805—1812), "*Regulations of the Friendly Society of Teachers in Fife and other Counties in Scotland instituted July 20th 1805*." (manuscript in library of IFoA).

Data availability statement. Data sharing not applicable – no new data generated.

Funding statement. No specific funding exists.

Competing interests. The authors declare none.

Cite this article: Wilkie AD (2025). Research that I have not yet done, but that I consider worth doing, *Annals of Actuarial Science*, **19**, 401–415. https://doi.org/10.1017/S1748499525100158

https://doi.org/10.1017/S1748499525100158 Published online by Cambridge University Press