ACTUARIAL RESEARCH CLEARING HOUSE
1990 VOL. 1

ON MODELLING SELECT MORTALITY

Iain D. Currie

Howard R. Waters

Iain D. Currie
Howard R. Waters
Department of Actuarial Mathematics and Statistics
Heriot-Watt University
Riccarton, Currie
Edinburgh EH14 4AS
Scotland

ABSTRACT

The paper is concerned with the graduation of mortality data and considers models that are functions of age and duration since selection. The methods are illustrated by applying them to the graduation of a data set relating to female assured lives in the United Kingdom.

1. Introduction

The Continuous Mortality Investigation Bureau (CMIB) is a research organisation established by the Faculty of Actuaries and the Institute of Actuaries in the UK. One of the main roles of the CMIB is to collect mortality data from UK insurance companies, to analyse these data, to issue regular reports on the analyses and, when required, to prepare new standard mortality tables. The most recent report of the CMIB, CMIR 9 [1988], discusses the analysis and graduation of several sets of data relating to the years 1979-82 inclusive.

The mortality of an 'assured life', i.e. an individual who has purchased a whole life assurance or endowment assurance policy, depends on, among other factors, the sex of the individual, his/her age and the duration since (s)he purchased the policy ('duration since selection' in the usual actuarial jargon). The traditional approach by the CMIB to the graduation of such data sets has been:

i) to graduate separately data from different sexes,
ii) to graduate separately, at least initially, data relating to durations 0-1 year, 1-2 years,..., 4-5 years and 5 or more years since selection, and then possibly combine some of the data for the higher durations so that standard tables with relatively short select periods can be produced.

The purpose of this paper, which is not part of the work of the CMIB, is to describe the graduation of one of the CMIB's data sets, modelling simultaneously age and duration since selection. The particular data set chosen related to the experience of female assured lives in the UK in the years 1979-82 inclusive. A full description of this data set and of the CMIB's graduations of it can be found in CMIR 9 [1988, §6].

The advantages of modelling simultaneously age and duration since selection are:

i) it uses data more efficiently by allowing us to infer information about ages/-durations where we have little data, from other ages/durations where we have more data,
ii) it makes it easier to prevent inconsistencies arising in the graduations; for example, if data for each year of duration since selection are graduated separately it is possible that at some ages, particularly at the ends of the age range being considered, graduated mortality rates may decrease with increasing duration since selection. Examples of this can be found in CMIR 9 [1988, Table 6.2]. This is a result of having little data at extreme ages rather than a real effect.
iii) it ought to give greater understanding of the nature and effect of selection.

Our approach to this problem is very much a statistical one; in other words, we have tried to find (and we hope succeeded in finding!) a function of age and duration since selection for the force of mortality which fits the data reasonably well, and also is consistent with our preconceived ideas (in particular that for a given attained age, the force of mortality should be a non-decreasing function of duration since selection). We have not attempted to develop a general 'law of mortality' as a

function of age and duration since selection, and for this reason the present paper is different in its approach from the paper by Tenenbein and Vanderhoof [1980].

In the following section we describe briefly the data available to us for this project. In §3 we discuss the problems arising from duplicate policies, i.e. individuals having more than one policy in the data. In §4 we describe the models of both the CMIB and ourselves and in §5 we describe the fitting of our models to the data. Finally, in §6 we give the results of our fits. This project has not yet been fully completed, and this paper could be regarded as a preliminary report. A fuller report on the project will be prepared and submitted for publication in the near future.

The authors are grateful to the CMIB for permission to use some of their data, and also grateful for the help given by their colleague, Professor John McCutcheon.

2. The Data

In this section we give some information about the data we have used for our graduations. A fuller description of these data and details of how they were collected can be found in CMIR 9 [1988], and earlier reports by the CMIB.

Our data relates to female assured lives (i.e. female policyholders with whole life or endowment assurance policies) in the years 1979-1982 inclusive. The data include lives who have not been medically examined as well as those who have, but only include lives accepted at normal premium rates.

The form of the data available to us is indicated in Table 1. For convenience, we have summed the data into 5-year age groups; the data supplied by the CMIB, and used by the authors, were in single years of age. For these data, 'age' means 'attained age nearest birthday' (not 'age when the policy was purchased'); 'duration' means 'duration in years since the policy was purchased'. Duration increases in one-year intervals up to 5 years; data for durations in excess of 5 years have been grouped. This feature caused us some problems which we will come back to later in this section. For each age (group) $\times$ duration cell we have two observed values: the number of deaths in the 4 year observation period (abbreviated Dths), and the central exposure during the observation period (abbreviated Exp). Data were available for ages above 90 but were ignored as they were considered by the CMIB to be somewhat unreliable.

Since our purpose was to model duration, it was inconvenient to have the data grouped for durations in excess of 5 years, especially as a large proportion of the total data was for durations in excess of 5 years. A preliminary task was to estimate, for each individual age, the average duration for policies where the duration was in excess of 5 years. The details of this estimation procedure will be given in the fuller report on this project to be prepared and submitted for publication in the near future. Table 2 shows, for selected individual ages, the estimated average duration for policies where the duration was in excess of 5 years.

3. Duplicate Policies

One of the problems with our data set is the presence of duplicate policies, i.e. two or more policies on the same life. Briefly, duplicates are a problem because their presence means it is not possible to assume that policies behave independently of each other. The CMIB conducted an investigation into the numbers of duplicate policies in this data set. This investigation was conducted by looking at the death certificates for a large proportion (56%) of the recorded deaths. If key information on two or more death certificates was identical, then it was assumed that the certificates related to the same life. This provided an estimate of the distribution of the number of policies per each individual life at each integer age among those individuals who died during the investigation period. It was then assumed that the distribution of the number of policies per each individual life at each age for all policyholders was the same as the distribution at the corresponding age among the deaths. The results of this investigation were made available to us by the CMIB and showed that there was little evidence of duplicate policies for durations up to 5 years, but some evidence of duplicates in the data for durations in excess of 5 years. In accordance with these findings we have made allowance for the presence of duplicate policies only in the data for durations in excess of 5 years.

Table 3 shows the estimates of the average number of policies per life for selected integer ages, and also the so-called 'variance inflation factor'. The variance inflation factor is defined to be

$$\sum_{i=1}^{\infty} i^2 n_i / \sum_{i=1}^{\infty} i n_i ,$$

where n_i is the number of individuals (for this particular age) who have exactly i policies. The role of the variance inflation factor is explained fully in Forfar, McCutcheon and Wilkie [1988] and will be mentioned briefly in the next section. (Strictly, the figures in Table 3 relate to durations in excess of 2 years, even though we applied them to the data for durations in excess of 5 years. This is because the corresponding data for durations in excess of 5 years were not available to us until recently.)

4. The Graduations

A full description of the methods of graduation used by the CMIB is given in Forfar, McCutcheon and Wilkie [1988] and their results are reported in CMIR 9 [1988]. In their graduations, the CMIB considered two closely related classes of functions for smoothing the raw death rates. We mention one of these classes, a generalisation of the Gompertz-Makeham mortality law, defined by

$$\mu_x = p_1(x) + \exp(p_2(x))$$

where μ_x is the force of mortality at age x and $p_1(x)$ and $p_2(x)$ are polynomials in x of orders r and s respectively (the order of a polynomial equals 1 plus its degree).

The CMIB refer to this as the Gompertz-Makeham formula of type (r, s), and write it as GM(r, s). For example, the GM$(2, 2)$ formula gives

$$\mu_x = a_0 + a_1 x + \exp(b_0 + b_1 x).$$

The function GM(r, s) is used to describe the underlying systematic relationship between the force of mortality, μ_x, and age, x. The observed variation about GM(r, s) is accounted for by a Poisson model which we describe next.

Let A_x denote the number of deaths between exact ages x and $x + 1$, and let R_x denote the central exposed to risk at age x. Then, assuming that the force of mortality is constant over each interval $[x, x + 1)$, we can suppose that A_x follows a Poisson distribution with mean $R_x \mu_x$; we write $A_x \sim \mathrm{P}(R_x \mu_x)$. We can summarise the approach of the CMIB in the following way:

$$A_x \sim \mathrm{P}(R_x \mu_x), \text{ where } \mu_x = \mathrm{GM}(r, s) \tag{4.1}$$

for some values of r and s.

With this model, the CMIB used three different fitting criteria to estimate the parameters in the functions GM(r, s)

(1) maximum likelihood,
(2) normal approximation to maximum likelihood,
(3) minimum χ^2.

It should be noted that (1) is not available if there are duplicates in the data, when only (2) and (3) can be used. CMIR 9 [1988] contains a very full description of the results of fitting the above models. In particular, there is much discussion of the values of r and s to be used. It should be recalled that the CMIB investigated each duration separately; thus, separate graduations were performed for each of the six durations 0-1,1-2,...,4-5,5+. Some durations were combined when it was decided that the mortality experiences of these durations were not significantly different. Generally, but not in all cases, the CMIB found that a GM(2,2) function gave satisfactory fits to the data. The full report is in §6 of CMIR 9 [1988].

The approach of the present paper is an extension of (4.1). The reasons for modelling age and duration simultaneously have already been discussed in §1 above. We experimented with a number of different models, and will report more fully on some of these other models at a later date. In the present paper we give details of the model that we finally settled on.

We shall denote by $\mu_{x,d}$ the force of mortality at exact attained age x and exact duration since selection d, both measured in years. We wanted our model for $\mu_{x,d}$ to be of the following form:

$$\mu_{x,d} = f_1(x) \times f_2(x, d) \times f_3(d)$$

where the functions f_1, f_2 and f_3 satisfied the following conditions:

(1) the age term, $f_1(x)$, was of the form GM$(0, s)$

(2) the interaction term, $f_2(x,d)$, was of a simple form,
(3) the duration term $f_3(d)$ was monotonic increasing and tended to a limit as $d \to \infty$.

The model that was finally selected was a modification of:

$$\begin{aligned}&(1)\ f_1(x) = \exp(a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4), \text{ i.e. GM(0,5)}\\ &(2)\ f_2(x,d) = \exp(a_5 x d)\\ &(3)\ f_3(d) = 1 - \exp(b_0 + b_1 d + b_2 d^2).\end{aligned} \tag{4.2}$$

The quadratic term in f_3 was used since it described (a) the large increase in mortality that is observed in moving from durations 0-1 to 1-2 and (b) the slow increase in mortality that is observed thereafter. A modification was required since at high durations the graduated mortality had a turning point corresponding, approximately, to the turning point of the quadratic function. We took the view that the turning point could be interpreted as the duration at which the effect of selection finally wore off. We adopted the simple expedient of flattening the graduated mortality rates. It should be noted that this flattening process was age dependent, i.e. we assumed that the effect of duration at fixed age x wore off at the turning value of $f_2(x,d)\ f_3(d)$. We make the following definition.

Definition For fixed x, $d_{\max}(x)$ is that value of d that maximises

$$\exp(a_5 x d) \times (1 - \exp(b_0 + b_1 d + b_2 d^2)).$$

The function $g(x) = d_{\max}(x)$ is of interest in its own right since it gives an indication of the length of the select period. A graph of $g(x)$ is provided in Figure 1 and an approximation to $g(x)$ is given by

$$g(x) = d_{\max}(x) = 12{\cdot}77 - 0{\cdot}562x + 0{\cdot}0107x^2.$$

We see clearly that the effect of duration wears off more quickly at lower ages. For example, at age 20 the selection effect has lasted for about six years, rising steadily to about ten years at age 60. Above age 60 the effect of selection becomes much more dramatic. For example, the force of mortality at age 70 is estimated as 0·0068 at duration zero, and 0·0164 at duration 20.

5. Fitting the Models

There are two difficulties to be coped with in fitting the models. The first difficulty is that the models of the CMIB and ourselves are non-linear. This is a computational problem only. The second difficulty is the problem of dealing with duplicate policies. The CMIB offered two solutions to the latter problem: the first solution was to use the normal approximation to the Poisson distribution for the number of deaths; their second solution was to use minimum χ^2. We took a third option and used a weighted Poisson. In the ordinary Poisson case, if $X \sim \mathrm{P}(\lambda)$ then $f(x) = e^{-\lambda}\lambda^x/x!$. The log likelihood is then

$$\begin{aligned}\ell(\lambda) &= x \log \lambda - \lambda + c\\ &= x\theta - \exp(\theta) + c, \quad \text{where } \theta = \log \lambda.\end{aligned}$$

The weighted Poisson is a random variable with log likelihood

$$\ell(\lambda) = w(x\theta - \exp(\theta)).$$

Clearly, if $w = 1$ we have an ordinary Poisson variable, while if $w \neq 1$ we have a variable with mean λ and variance λ/w. The appropriate weighting factor is $1/r_x$ where r_x is the variance inflation factor. If A_x, R_x and r_x are the number of claims, the number of policy years and the variance inflation factor respectively, then the contribution to the log likelihood at age x is

$$\ell(\mu_x) = (A_x \log \mu_x - R_x\mu_x)/r_x'.$$

This gives $\hat{\mu}_x = A_x/R_x$ and estimated variance $r_x A_x/R_x^2$. Asymptotically, this is the same solution as obtained using a normal approximation or a minimum χ^2 approach; if $w = 1$ the exact log likelihood is used.

The computations were carried out using the statistical programming language Genstat 5 [Genstat 5 Committee, 1987]. Genstat 5 is a Fortran based statistical command language which is particularly suitable for fitting models of the forms described in §4; a high quality graphics interface allows the production of the graphs presented in the next section.

6. Results

The model (4.2) with a weighted Poisson likelihood was fitted to the data with ages 20 through 90 and durations 0-1,1-2,2-3,3-4,4-5,5+; duration 5+ used estimated mean durations for each age. The fitted parameter values, together with standard errors and t-values are given in Table 4.

The fitted model gives a deviance (-2×log likelihood ratio) of 503 with about 420 degrees of freedom and thus a mean deviance of 1·2 (the exact number of degrees of freedom are unknown because of the flattened quadratic). At first sight, this figure may seem rather high (if the model is exactly true the mean deviance should be about 1). However, examination of the residuals does not reveal any obvious departure from the model; it is more that the residuals are generally rather large. A possible explanation is that there were rather more duplicates in the data than were allowed for; this would certainly have the effect of inflating the residual variance.

Figure 2 shows $\mu_{x,d}$ as a function of x for three values of d. In other words it shows the force of mortality as a function of attained age for durations 0, 4 and 10 years. It can be seen from Figure 2 that for fixed x we have

$$\mu_{x,0} < \mu_{x,4} < \mu_{x,10}$$

and that the relative difference between $\mu_{x,4}$ and $\mu_{x,0}$ is greater than that between $\mu_{x,10}$ and $\mu_{x,4}$. These features are not surprising since we would expect the effect

of duration since selection to decrease with increasing duration (and have chosen our graduating function for $\mu_{x,d}$ accordingly!). This feature is shown clearly again in Figure 3, which gives the graphs of $\mu_{x,d}$ as a function of duration d for various attained ages x. Our 'flattening' of the duration effect can be seen in Figure 3 for attained ages 30, 50 and 60. For attained age 70 the effect of selection has lasted for about 30 years (see also Figure 1) and so the graph of $\mu_{70,d}$ in Figure 3 is increasing for $0 \leq d \leq 25$.

Figure 4 is also a graph of the force of mortality as a function of duration but differs from Figure 3 in that it shows a graph of $\mu_{x_0+d,d}$ as a function of d, where x_0 can be regarded as the age at which the policy was effected. Hence Figure 4 shows, for an individual who purchased her policy at age 30 (or age 45 or age 60), the development of the individual's force of mortality as her age and duration increase simultaneously.

This paper has given a very brief description of a project which is nearing completion. As we have mentioned several times in this paper, we intend to prepare for publication a fuller report on this project in the near future. The fuller report will not only explain in more detail points which have been mentioned in this paper (e.g. the estimation of the average duration of a policy whose duration is in excess of 5 years) but will also:

a) give details of the other methods tried by us for modelling age and duration simultaneously,
b) consider the actuarial implications of the graduations (e.g. the implications for premium rates),
c) compare our graduations with those of the CMIB, and, possibly,
d) give details of the application of the techniques of this paper to other data sets.

References

CMIR 9 [1988] Report Number 9 of the Continuous Mortality Investigation Bureau. Faculty of Actuaries, Edinburgh, and Institute of Actuaries, London.

Forfar, D.O., McCutcheon, J.J. and Wilkie, A.D. [1988] On graduation by mathematical formula. J.Inst.Act. **115**, part I, 1-149.

Genstat 5 Committee. [1987] Genstat 5: Reference Manual. Clarendon Press, Oxford.

Tenenbein, A. and Vanderhoof, I.T. [1980] New mathematical laws of select and ultimate mortality. Trans.Soc.Act. **XXXII**, 119-158.

Table 1 The data summed into 5-year age groups

Age	Duration					
	0–1		1–2		2–3	
	Exp	Dths	Exp	Dths	Exp	Dths
20-24	120276	18	105301	30	82547	17
25-29	116589	29	109487	28	96635	31
30-34	120652	28	114904	34	102733	31
35-39	96696	39	93893	56	86006	45
40-44	73863	37	71023	53	65275	46
45-49	59871	49	57706	75	53902	79
50-54	43419	88	44167	99	42961	102
55-59	21001	47	22126	90	23070	108
60-64	9640	27	8938	46	8764	57
65-69	4190	20	3877	39	3583	34
70-74	1812	22	1495	27	1269	11
75-79	480	1	558	13	499	5
80-84	71	1	84	1	91	1
85-90	6	0	9	0	11	1
Total	668563	406	633565	591	567345	568
	3–4		4–5		5+	
	Exp	Dths	Exp	Dths	Exp	Dths
20-24	64394	13	44759	10	54323	16
25-29	84640	28	71921	24	187983	68
30-34	89745	28	76367	25	295413	128
35-39	76613	33	65334	39	290985	189
40-44	58937	50	50937	41	240867	266
45-49	48857	73	42825	73	228286	404
50-54	40297	95	35948	92	223775	756
55-59	23258	93	22653	103	194852	951
60-64	8860	59	9137	47	93376	653
65-69	3334	38	3057	25	27005	277
70-74	1150	6	1166	19	12111	258
75-79	448	6	390	10	6135	205
80-84	90	1	96	4	3250	226
85-90	15	0	15	1	1718	210
Total	500636	523	424602	513	1860077	4607

Table 2 Estimated average duration for policies where the duration is in excess of 5 years.

Age	Est. Dur.	Age	Est. Dur.
20	6·27	60	10·00
25	6·53	65	9·78
30	7·36	70	9·89
35	7·88	75	13·99
40	8·68	80	18·89
45	8·65	85	19·55
50	9·63	90	22·39
55	9·96		

Table 3 Estimates of the average number of policies per life and of the variance inflation factor.

Age	Average number of policies per life	Variance inflation factor
20	1·000	1·00
25	1·118	1·21
30	1·083	1·15
35	1·139	1·39
40	1·032	1·06
45	1·088	1·16
50	1·085	1·16
55	1·092	1·21
60	1·094	1·23
65	1·074	1·14
70	1·057	1·16
75	1·045	1·09
80	1·087	1·16
85	1·333	3·25
90	1·000	1·00

Table 4 Estimated coefficients, standard errors and t-values

	Coeff.	Standard error	t-value
a_0	-3·980	0·027	-145·1
a_1	3·942	0·168	23·5
a_2	0·022	0·398	0·1
a_3	4·287	1·136	3·8
a_4	4·512	0·894	5·0
a_5	0·045	0·010	4·6
b_0	-0·449	0·041	-10·9
b_1	-0·199	0·028	-7·2
b_2	0·004	0·003	1·3

Note: The above coefficients are to be applied to (4.2) using the transformed age $x^* = (x - 70)/50$.

Figure 1 - The Function d-max(x)

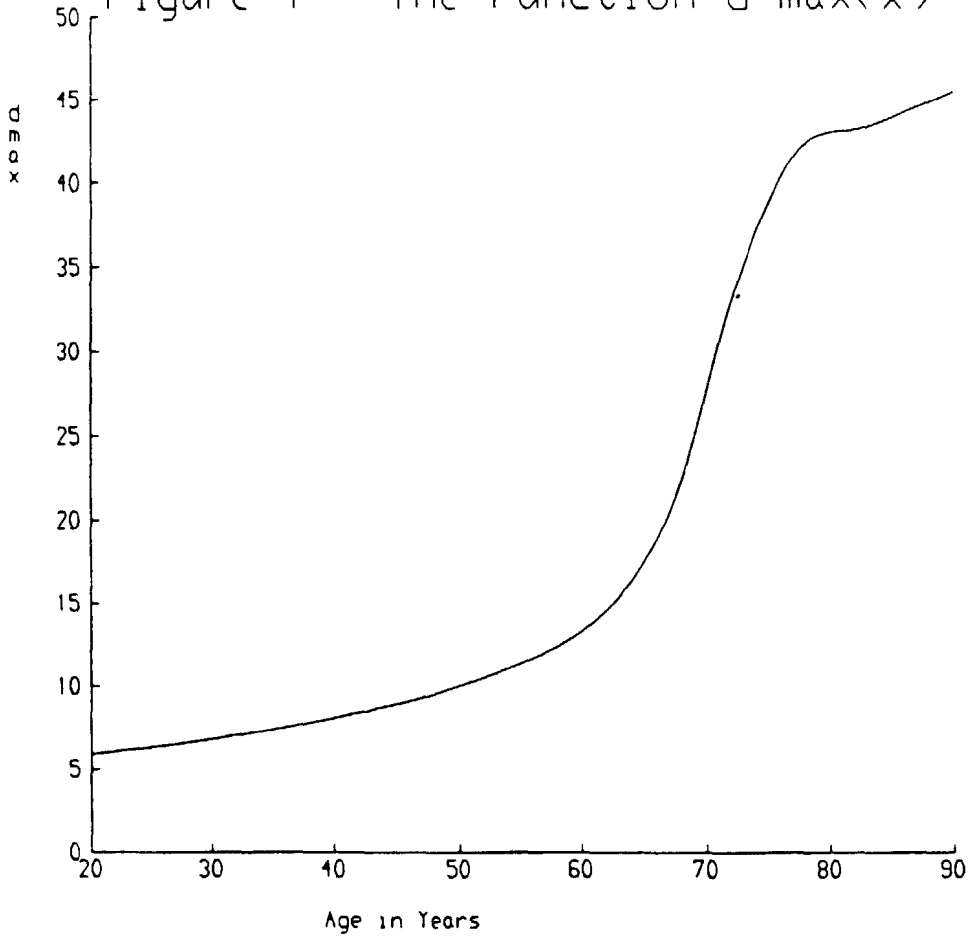

Figure 2 - Force of Mortality v Age

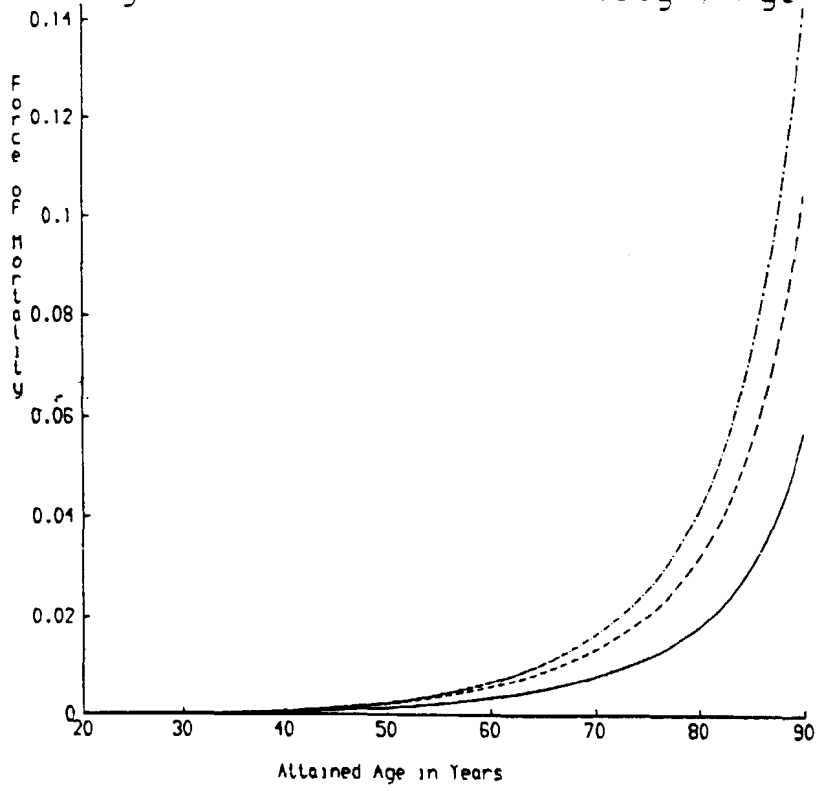

Figure 3 - Force of Mortality v Duration

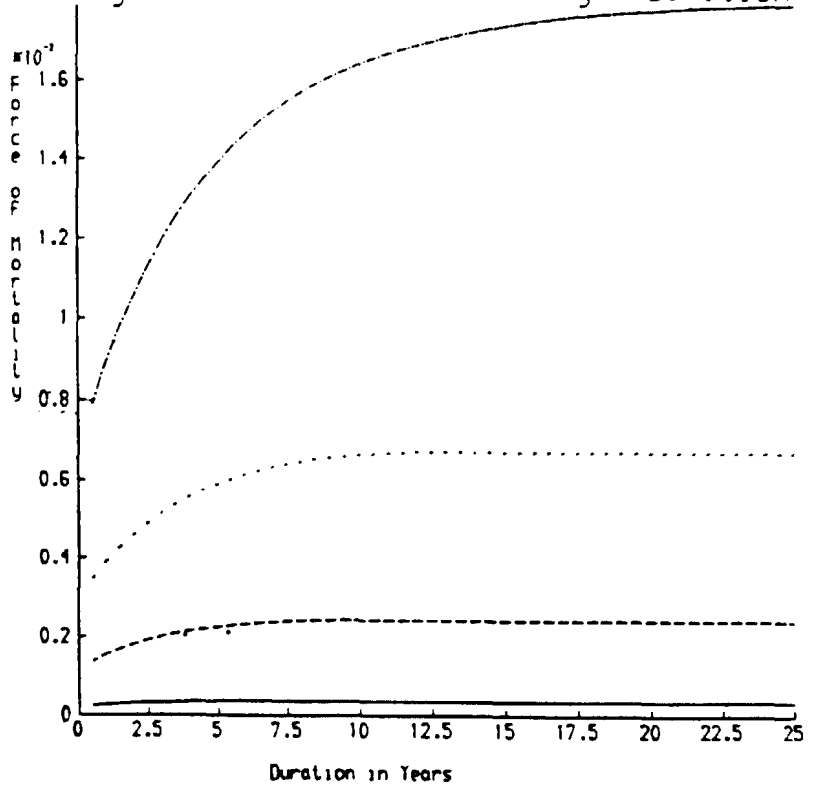

Figure 4 - Force of Mortality v Duration