Practical "Modern" Bayesian Statistics in Actuarial Science

Fernanda Pereira

PRACTICAL "MODERN" BAYESIAN STATISTICS IN ACTUARIAL SCIENCE

ABSTRACT

The aim of this paper is to convince actuaries that Bayesian statistics could be useful for solving practical problems. This affirmation is due to two main characteristics of Bayesian modelling not yet fully explored by practitioner actuaries: first, the possibility of absorbing subjective information and second, the wider range of models available in the Bayesian framework.

In order to help in this "convincing" process this paper includes an overview of Bayesian statistics in actuarial science and cites many published papers based on this theory with insurance applications. An approach with as few formulae as possible will be used to make it easier to follow for all actuaries, independently of their involvement in statistics.

ACKNOWLEDGEMENTS

I acknowledge the suggestions made by my Ph.D. supervisor, Dr. Richard Verrall, through the many reviews which he did. I also am grateful to the financial support of Funenseg (National School Foundation in Insurance), Rio de Janeiro (BR), which gave me a grant to follow the Ph.D. program in actuarial science at City University, London (UK).

PRACTICAL "MODERN" BAYESIAN STATISTICS IN ACTUARIAL SCIENCE

1 INTRODUCTION

" (...) within the realm of actuarial science there are a number of problems that are particularly suited for Bayesian analysis."
Klugman (1992)

Bayesian theory is a powerful branch of statistics not yet fully explored by practitioner actuaries. One of its main benefits, which is the core of its philosophy, is the ability of including subjective information in a formal framework. Apart from this, the wide range of models presented by this branch of statistics is also one of the main reasons why it has been so much studied recently. Artificial intelligence and neural networks are examples of new disciplines that are heavily based on Bayesian theory.

Bayesian theory has been one of the most discussed and developed branches of statistics over the last decade. There have been an enormous number of papers published by a large number of statistical researchers and practitioners. The recent developments are mainly due to, firstly, the recent computer developments that have made it easier to performs calculation by simulations and, secondly, to the failure of classical statistic methods to give solutions to many problems.

But, although so many developments have been occurring in Bayesian statistics very few actuaries are aware of them and even fewer make use of them. Throughout the works reviewed in this paper it is possible to observe that the authors believe that Bayesian statistics can add great value to the role of a practitioner actuary and, in a way, this paper forms a kind of manifesto.

Since the advent of credibility theory, which has at its core Bayesian statistics, this statistical philosophy has not been greatly exploited by practitioner actuaries. It was in 1914 that the first paper on credibility theory was published. This theory made actuaries one of the first practitioners to use the Bayesian philosophy. Since then many developments in credibility theory have occurred, but it is probably the only tool based on Bayesian theory used in an office environment, and even this is rare. However, judgement is used in an everyday basis and it is often argued that in this way an informal Bayesian approach is used.

This paper expounds the development of Bayesian models in actuarial science in academia, but of which, it is believed, very few practitioners are aware. In this way the paper aims to build a bridge between modern Bayesian statistics and practical problems. In the following sections around 20 models described in published papers in actuarial journals will be rewritten in a more informal way, avoiding the extensive calculations normally needed in a Bayesian application. On one hand it means that actuaries that do not have a deep involvement in statistics can understand the ideas behind the models. On the other hand it will not be possible to fully explain all the calculations behind the models. So it should be stressed that the more interested reader is encouraged to refer to the original papers in order to get a deeper explanation

of the respective models.

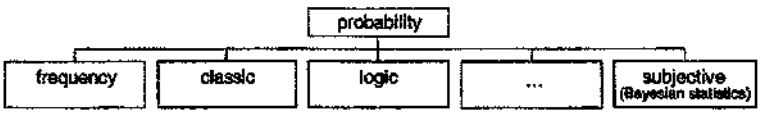
The outline of the paper is as follows. In section 2, an introduction to Bayesian theory is presented. Section 3 explains Bayesian approaches to traditional methods such as credibility theory, claims reserving and graduation. These models are discussed again in section 4 after an introduction to simulation has been given. In section 5 models entirely built in a Bayesian framework are presented. Section 6 contains the conclusions.

2 INTRODUCTION TO BAYESIAN THEORY

" (...) all values are still selected on the basis of judgement, and the only demonstration they (*actuaries*) can make is that, in actual practice, it works. (...) It does work!"
Bailey (1950)

As is well known, probability theory is the foundation for statistics. The differences in the interpretation of the term probability define also the respective differences in statistical theories. As examples, there are probability theories based on frequency, classical, logic and subjective philosophies. The last one is the core of Bayesian statistics.

Figure 1



The subjective interpretation states that the probability that an analyst assigns to a possible outcome of a certain experiment represents his own judgement of the likelihood that a specific outcome will be obtained. This judgement will be based on the analyst's beliefs and information about the experiment. As a contrast, frequency statistics, for example, do not include formally this judgement but only the information received from the observation set itself.

Bringing those interpretations to the inference problem of estimating a specific parameter, Bayesian statistics differs clearly from the others. In classical and frequency statistics the analyst is searching for a best estimator of a parameter that has a true value, but which is unknown by him. In the Bayesian statistics the analyst does not believe in this true value, but in a range represented by the previous information that he has.

The recognition of the subjective interpretation of probability has the salutary effect of emphasising some of the subjective aspects of science. It also defines a formal way of including judgement about the process in the chosen model. This subjective information is included in the model by defining a prior distribution for the

unknown parameters.

Bayes theorem is the formal mechanism of incorporating prior information into the modelling. This theorem mixes the prior subjective information with that observed in the experiment, producing a posterior distribution. This distribution is considered as an update of the previous judgement (prior) through the data observed (likelihood).

More formally Bayes theorem is defined as follows. Consider a process in which observations (X is the vector of observations[1]) are to be taken from a distribution for which the probability density function is $p(X|\theta)$, where θ is a set of unknown parameters. Before any observation is made, the analyst would include all his previous information and judgements of θ in a prior distribution $p(\theta)$, that would be combined with the observations to give a posterior distribution $p(\theta|X)$ in the following way:

$$p(\theta|X) \propto p(X|\theta)p(\theta) \qquad (1)$$

It is only after the posterior distribution is fully defined, that the estimation is performed. It means that only after having all information at hand that the analyst will define which estimation will be used. This definition step is called "decision theory". So, if the analyst searches for a specific value, called "point estimate", he can consider the mean, mode or any other statistics as the estimator (which depends on the chosen loss function). A range, like a confidence interval, can also be calculated. A further explanation of Bayesian theory can be found in any of the references on this theory listed at the end of the paper.

In order to illustrate Bayesian statistics and show the difference of approaches among statistical theories, a numerical problem is presented. This is a very simple example and it was chosen in order to show step by step the concept behind Bayesian analysis.

Consider a policyholder whose claim values (which are independent and identical distributed) come from a specific distribution $p(x|\theta)$. Suppose we have observed annual claims for 5 years with the following results:

124.93 110.67 106.93 104.05 101.60

Using the well known linear model, it would state that such values are from a normal distribution with unknown mean θ and known variance σ^2 ($x_i \sim$ normal (θ, σ^2) $i=1,\ldots,5$). In this model the maximum likelihood estimator (MLE) of θ is equal to the sample mean:

$$\bar{x} = 109.64 \qquad (2)$$

[1] In this paper upper caps stand for vector and matrix, and small caps for single values

In the Bayesian solution the approach would be different. Now the unknown parameter θ is also considered as a random variable and this interpretation is formalised by the inclusion of a prior distribution. The choice of this distribution is completely the analyst's subjective decision. In the example described here we could, for instance, look for another policyholder with similar characteristics or even some other previous experience.

Suppose that this investigation suggested the value of 100 for θ. In this case the following prior distribution is chosen[2]:

$$\theta \sim \text{normal}\,(100\,, \sigma^2_{\text{prior}}) \tag{3}$$

with σ^2_{prior} known, as suitable. Now, using Bayes theorem, the Bayesian minimum least square estimator of θ is given by the formula[3]:

$$\frac{5\bar{x}\sigma^2_{prior} + 100\sigma^2}{5\sigma^2_{prior} + \sigma^2} \tag{4}$$

with 5 as the sample size.

It is straightforward to observe that the formula (4) is more complicated than the one derived in (2). Formula (4) also includes the variances in order to calculate the mean estimator, where the Bayesian estimator is a weighted mixture of the prior and sample information. This mixture is clearer when the formula (4) is rewritten as follows:

$$\bar{x} * z + 100 * (1 - z) \tag{5}$$

$$\text{with } z = \frac{5\sigma^2_{prior}}{5\sigma^2_{prior} + \sigma^2}$$

Now it is possible to see that if instead of 5 an infinitely large size for the sample were given, all of the weight would go to the sample mean ($z = 1$), giving the solution in formula (2). On the other hand, if the value for σ^2 were infinitely large, the weight would go to the prior distribution mean ($z = 0$).

Proceeding with the analysis of the Bayesian model, it would be interesting to consider the mixture that is taking place. In order to do this, the full posterior distribution $p(\theta|X)$ will be defined. This distribution is, again, a normal distribution with mean given by (4) and variance:

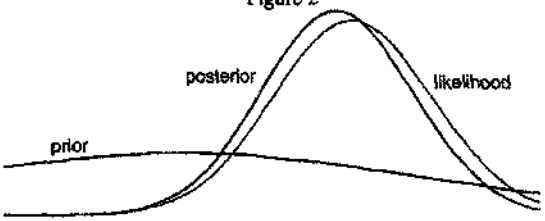
$$\frac{\sigma^2_{prior}\sigma^2}{5\sigma^2_{prior} + \sigma^2} \tag{6}$$

---

[2] Normal distribution was conveniently chosen by conjugacy. See references for further explanation.
[3] This calculation is found in any reference of Bayesian theory at the end of the paper.

Now it is necessary to fix the variance parameters. Taking a subjective approach, define σ^2 to be 100 and σ^2_{prior} to be 200. This gives a posterior mean and variance of 108.76 and 18.18 respectively. The likelihood (when $p(X|\theta)$ is seen as a function of θ) will also be normally distributed with mean 109.64 from the data, and variance 20 (σ^2/(sample size)). The plot of the prior and posterior distributions with the likelihood is in figure 2.

Figure 2



This shows the way in which Bayes theorem allows the mixture of information. Observe that the posterior distribution is close to the likelihood but still keeping some influence from the prior.

It is argued that Bayesian theory gives a better description of what is going on, since it does not give just a point estimate, but also a distribution related to the parameter. But to apply it, many more calculations are needed to achieve this estimation, even in this simple example. When σ^2 is unknown, for instance, it is necessary also to define a prior distribution for this parameter and the calculations become even more complicated.

When a subjective approach is taken a prior distribution can be hard to define and even harder to justify. In fact, it is one of the most controversial elements in Bayesian statistics. If an analyst does not want to include prior information, but does want to use a Bayesian approach, a non-informative prior may be included. In figure 2 it would mean that the shape of the prior distribution would be completely flat and the posterior distribution would be the same as the likelihood (z =1).

There are many ways of defining a non-informative prior. The main objective is to give as little subjective information as possible. So, usually a prior distribution with a large value for the variance is used. Another way of including the minimal prior information is to find estimates of the parameters of the prior distribution, using the data. This last approach is called empirical Bayes, but often there is a relationship between those two approaches – non-informative and empirical Bayes – that will not be developed further here[4].

Theoretically, a prior distribution could be included for all the parameters that

[4] For further details see any of the references of Bayesian theory at the end of the paper.

are unknown in a model, so that any model could be represented in a Bayesian way. However, this often leads to intractable problems (mainly integrals without solution). So the main limitation of Bayesian theory is the difficulty, and in many cases the impossibility, of analytically solving the required equations.

In the last decade many simulation techniques have been developed in order to solve this problem and to obtain estimates of the posterior distribution. These techniques were turning points for the Bayesian theory, making it possible to apply many of its models. On one hand, the use of a final and closed formula for a solution is, generally speaking, more satisfactory than the use of an approximation through simulation. On the other hand, simulation gives a larger range of models for which solutions (or at least good approximations) can be obtained.

Now some more elaborate examples will be explored. In the next section models with analytical solutions will be presented, and in the section 4 the same problems will be reanalysed using a simulation approach. The models used in section 3 and 4 are listed in the following table, split by subject, type of solution and data set used:

Paper	Subject[5] (section)	Type[6]	Data
Bühlmann and Straub (1970)	CT (3.1)	ANA	Klugman (1992)
Klugman (1992)	CT (3.1)	APP	Klugman (1992)
Verrall (1990)	CL (3.2)	ANA	Taylor and Ashe (1983)
Klugman (1992)	GR (3.3)	APP	Klugman (1992)
Kimeldorf and Jones (1967)	GR (3.3)	ANA	London (1985)
Pereira (1998)	CT (4.1)	SIM	Klugman (1992)
Charissi (1997)	CL (4.2)	SIM	Taylor and Ashe (1983)
Ntzoufras and Dellaportas (1997)	CL (4.2)	SIM	Ntzoufras and Dellaportas (1997)
Kouyoumoutzis (1998)	GR (4.3)	SIM	Kouyoumoutzis (1998)
Carlin (1992)	GR (4.3)	SIM	Carlin (1992)

3 TRADITIONAL METHODS

"Statistical methods with a Bayesian flavour (...) have
long been used in the insurance industry (...)."
Smith et all (1996)

This section looks at some traditional areas of actuary theory: credibility theory, the chain ladder model and graduation. Apart from one example cited in the credibility theory subsection, all the models used in the following subsections have an analytical solution. They are then used as illustrations of where modern Bayesian theory can be applied without any approximation.

[5] CT = Credibility Theory; CL = Chain Ladder; GR = Graduation.
[6] ANA = Analytical; APP = Approximation, but not simulation; SIM = Simulation.

The section is divided as follows. Subsection 3.1 is about credibility theory and two approaches will be described: one which was originally used when credibility theory was introduced, and a purely Bayesian approach. The chain ladder technique and graduation are also reviewed in a Bayesian framework in subsections 3.2 and 3.3.

3.1 CREDIBILITY THEORY

Credibility theory was first introduced in 1914 by a group of American actuaries almost at the same time as the Casualty Actuarial Society was created. At that time those actuaries had to define a premium for a new insurance product – "workmen's" compensation – so they based the tariff on a previous kind of insurance which was substituted by this one.

As new experience arrived, a way of including this information was formalised, mixing the new and the old experiences. This mixture is the basis of credibility theory, which searches for an credibility estimator that balances the new but volatile data, and the old but with a historical support. Most of the research until 1967 went in this direction, creating the branch of credibility theory called limited fluctuation.

The turning point in this theory, and the reason why it is used nowadays, happened when actuaries realised that they could bring such a mixture idea inside a portfolio. This new branch searches for an individual estimator (or a class estimator), but still using the experience for the whole portfolio. Such an estimator would consider the "own" experience on one side, but giving more confidence to it by also including a more "general" one on the other side. In a way, it formalises the mutuality behind insurance, without the loss of the individual experience.

There are many papers discussing this theory, but the one by Bühlmann (1967) is generally seen as a landmark. In this paper credibility theory was completely formalised, giving a basic formula and philosophy. Since then, many models have been developed. Given that credibility theory is completely based on Bayesian statistics, a bibliographical review is presented at the end of this paper.

In order to illustrate credibility theory the Bühlmann and Straub (1970) model is used, which is a step forward from Bühlmann (1967)[7]. The data set is taken from Klugman (1992), which is the first book on Bayesian statistics in actuarial science. The observations are the number of claims (y_{ij}) for 133 occupations (i=1,...,133) in workers' compensation insurance with 7 years experience (j=1,...,7). The respective amount of the payroll (w_{ij}) is also known and is used as a weight for each occupational class. In order to explain the data the history for class 11 is given in the following table:

[7] All formulae for both models are given in appendix A.

Class	Year	Payroll	y_{ij}
11	1	149.683	6
11	2	157.947	6
11	3	174.549	5
11	4	181.317	10
11	5	202.066	13
11	6	187.564	7
11	7	229.830	8

Modelling the frequency ratio x_{ij} (y_{ij}/w_{ij}) by the Bühlmann and Straub model gives the following distributions:

$$\begin{aligned} x_{ij} \mid \theta_i &\sim \text{normal}(\theta_i,\ \sigma^2/w_{ij}) \\ \theta_i &\sim \text{normal}(\mu, \tau^2) \end{aligned} \qquad (7)$$

For all i and j, and σ^2, μ and τ^2 known. Now, with $\bar{x}_i$ as the observed mean and z_i as the credibility factor for class i, the credibility estimator for the class ratio θ_i is:

$$\bar{x}_i \times z_i + \mu \times (1 - z_i) \qquad (8)$$

The solution proposed by Bühlmann and Straub is to calculate the values of σ^2, μ and τ^2 from the observations, substituting these values and coming out with the solution for the formula above. Proceeding with their calculation changes formula (8) to:

$$\bar{x}_i \times \hat{z}_i + \bar{\bar{x}} \times (1 - \hat{z}_i) \qquad (9)$$

where $\hat{z}_i$ is the estimated value of z_i, after including the values for the variances and $\bar{\bar{x}}$ is the overall observed mean.

It may not be clear where the prior information has been inserted into this model. The reason for this is that the formula (9) was developed in order to balance the information of the class own observed experience, $\bar{x}_i$, with the observed overall one, $\bar{\bar{x}}$. In this model the distribution $p(\theta)$ is not playing a role of a real Bayesian prior, but its parameters are substituted by the values calculated on the data set.

This type of solution is called empirical Bayes approach. In order to have a fully subjective Bayesian solution another level of distribution would have to be included. This would contain information about the parameters σ^2, μ and τ^2, which are considered unknown. In this way the model in (7) would include three levels and be changed to[8]:

$$\begin{aligned} x_{ij} \mid \theta_i, \sigma &\sim \text{normal}(\theta_i,\ \sigma^2/w_{ij}) \\ \theta_i \mid \mu, \tau &\sim \text{normal}(\mu, \tau^2) \\ p(\sigma, \mu, \tau) & \qquad (\text{for all } i \text{ and } j) \end{aligned} \qquad (10)$$

[8] Observe that once σ, μ and τ are considered unknown $p(\theta) \neq p(\theta \mid \mu, \tau)$ for instance.

In this new model $p(x_{ij}|\theta_i, \sigma)$ stands for each class experience, $p(\theta_i|\ \mu, \tau)$ for the overall portfolio information and $p(\sigma,\mu,\tau)$ brings the prior distribution for the unknown parameters in the previous distributions. Now, $p(\theta_i|\ \mu, \tau)$ informs only that each class mean comes from the same distribution. Unfortunately unless very strong assumptions for $p(\sigma,\mu,\tau)$ are included, it is not possible to derive the posterior distributions for θ_i.

In order to use a pure Bayesian approach, Klugman (1992) included priors for σ^2, μ and τ^2, but in a "non-informative" way. No analytical solution is available and an approximation technique (Gaussian quadrature) was used. Both solutions[9] are shown for some classes in the following table:

	Solution				Forecasting			
Class	$\sum_{j=1}^{6} w_{ij}$	$\bar{x}_i$	Bühlmann and Straub	Klugman	w_{i7}	y_{i7}	Bühlmann and Straub	Klugman
4	0.037	0.0	0.03949	0.04045	-	0	-	-
11	1,053.126	0.04446	0.04345	0.04422	229.83	8	9.99	10.16
112	93,383.54	0.00188	0.00201	0.00193	18,809.7	45	37.81	36.30
70	287.911	0.0	0.02059	0.01142	54.81	0	1.13	0.63
20	11,075.31	0.03142	0.03164	0.03151	1,315.37	22	41.62	41.45
89	620.968	0.42997	0.29896	0.36969	79.63	40	23.81	29.44
					Forecast error[10]		15.55	13.20

However, it is sometimes desirable to include more information in $p(\sigma,\mu,\tau)$, which would also mean more difficulty in calculating the solution. In order to overcome such problems powerful simulation techniques have been developed in recent years and subsection 4.1 shows how to apply them.

3.2 CLAIMS RESERVING

Claims reserving is one of the most important branches in the general insurance area of actuarial science. Usually a macro model, where data are accumulated by underwriting year and development year, is used, and the data are given in a triangular format. One of the features of those models is the small amount of data available for the later development years, and this gives a large degree of instability to any estimate. Actuaries overcome this problem through professional judgement when they choose factors or consider benchmarks.

---

[9] Although data were observed for 7 years the two solutions only use 6 years to do the calculations.

[10] Forecast error $= \sum_{i=1}^{133} (Forecast - Y_{i7})^2 / w_{i7}$

Here another way of including this subjective information will be given, which is more formal, statistically speaking, since it uses a prior distribution. The approach used here is the chain ladder technique, which is one of the most popular macro methods to predict claims reserves. But in the following examples no inclusion of the tail factor will be considered.

The data comes from Taylor and Ashe (1983), and the exposure factor per underwriting years and the data are given below, where the influence of the exposure has to be taken out from the claim amount before any analysis.

Exposure: 610 721 697 621 600 552 543 503 525 420

Underwriting year \ Development year										
	357848	766940	610542	482940	527326	574398	146342	139950	227229	67948
	352118	884021	933894	1183289	445745	320996	527804	266172	425046	
	290507	1001799	926219	1016654	750816	146923	495992	280405		
	310608	1108250	776189	1562400	272482	352053	206286			
	443160	693190	991983	769488	504851	470639				
	396132	937085	847498	805037	705960					
	440832	847631	1131398	1063269						
	359480	1061648	1443370							
	376686	986608								
	344014									

In Kremer (1982), which is a paper on credibility theory, the chain ladder is proved to be similar to the two way analysis of variance linear model expressed by:

$$x_{ij} = \ln(y_{ij}) \qquad (11)$$

With x_{ij} independent normal(θ_{ij}, σ^2), where $\theta_{ij} = \mu + \alpha_i + \beta_j$ and y_{ij} as the incremental value of the claims for row (underwriting year) i and column (development year) j.

The solution of Kremer (1982) is to calculate the MLE of the unknown parameters together with the estimate of σ^2. In Verrall (1990), which is the paper reviewed here, the same model is used but a Bayesian solution is applied. In fact three Bayesian solutions are presented: "pure Bayes without prior information", "pure Bayes with prior information" and "empirical Bayes". The formulae of those models are given in full in appendix B, but here the main ideas behind each model are given. It is interesting to notice that in order to have an analytical solution, none of these models includes a prior distribution for the variance parameters.

Proceeding the explanation, a prior distribution is attached to the model in (11) that is rewritten in a matrix notation:

$$\begin{aligned} X \mid \theta &\sim \text{normal}\,(K\theta,\ \sigma^2 I) \\ \theta \mid \theta_1 &\sim \text{normal}\,(\theta_1,\ \Sigma) \end{aligned} \qquad (12)$$

where $X=(x_{11},\ldots,x_{1n},x_{21},\ldots,x_{2n-1},\ldots,x_{n1})$,
$\theta=(\mu, \alpha_2,\ldots, \alpha_n, \beta_2, \ldots, \beta_n)$,
$\Sigma=\text{diag}(\sigma_\mu^{-2}, \sigma_\alpha^{-2},\ldots, \sigma_\alpha^{-2}, \sigma_\beta^{-2},\ldots, \sigma_\beta^{-2})$,
$\theta_1=(\mu^*, \alpha_2^*, \ldots, \alpha_n^*, \beta_2^*, \ldots, \beta_n^*)$,
K is the design matrix in order to produce the model in (11),
I as the respective identity matrix,
$\sigma^2, \sigma_\mu^{-2}, \sigma_\alpha^{-2}, \sigma_\beta^{-2}$ are known variances,
$\alpha_1=\beta_1=0$ for uniqueness (see Verrall (1990) for details).

The "pure Bayes without prior information" uses a non-informative prior approach. In this way, $\sigma_\mu^{-2}, \sigma_\alpha^{-2}$ and σ_β^{-2} go to zero and the model solution gives exactly the same results as the classical and usual MLE used in Kremer (1982).

But more information could be inserted straight into this second level distribution, instead of using the non-informative one for all parameters. This is the "pure Bayes with prior information" approach and will be applied by changing θ_1 and Σ in order to keep the non-informative approach for parameters $(\mu, \beta_2, \ldots, \beta_n)$, but not for the row parameters. Proper prior distributions for $(\alpha_2,\ldots,\alpha_n)$ are defined, but they are hard to define, since there is no intuitive explanation related to them. In this example the following set of prior distribution (based on the result obtained at the MLE model) was chosen:[11]

$$\alpha_i \sim \text{normal}\ (0.3,\ 0.05); \qquad (13)$$

for all $i = 2,\ldots,n$.

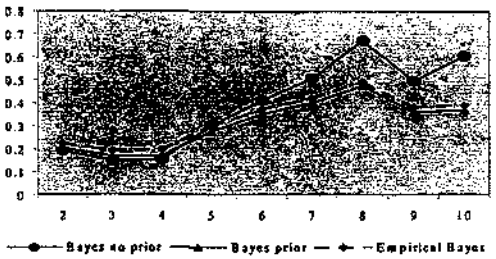
The third approach, "empirical Bayes" is based on the credibility theory assumption, that there is some dependency among the parameters related to the row and they are not really independent as before. So, in formula (12) the non-informative approach is kept for $(\mu, \beta_2, \ldots, \beta_n)$, ($\sigma_\mu^{-2}$ and $\sigma_\beta^{-2} = 0$), but a different one is imposed for the row parameters.

Now, instead of defining a distribution like (13), the general distribution (12) is kept and another level of prior distribution is added to $(\alpha_2^*, \ldots, \alpha_n^*)$, with a non-informative approach. In this way no prior value is given, but only a dependency among the row parameters is imposed.

All three models were applied to this data set. "Pure Bayes without prior information", which is the equivalent to the MLE solution by Kremer (1982), had the worse performance when compared to the other two in all analysis done by Verrall (1990). "Pure Bayes with prior information" and "empirical Bayes" also had a better smoothness to the row parameters as can be seen in figure 3:

[11] which is the same as $\alpha_i^* = 0.3$ for $i=2,..,n$ and $\sigma_\alpha^2=0.05$.

Figure 3



All these models have analytical solutions, but a prior distribution for the variance parameters was not used. The way in which this can be done will be explained in subsection 4.2.

3.3 GRADUATION

Graduation is an important part of the job of a life actuary and many methods have been developed in order to carry it out. Using the definition from Haberman (1996) "graduation may be regarded as the principles and methods by which a set of observed probabilities are adjusted in order to provide a suitable basis for inference to be drawn and further practical computations to be made".

The usual data set where graduation is applied includes the number of policyholders in the beginning of the observation period (usually one year) and at its end the number of occurred deaths is accounted. In order to illustrate it, the following sample was taken from London (1985):

Age (i)	Frequency rate (x_i)	Number of:	
		Policyholders (n_i)	Deaths (d_i)
63	0.00928	9,487	88
64	0.01226	10,770	132
65	0.01100	24,267	267
66	0.01120	26,791	300
67	0.01481	29,174	432

Whittaker graduation is one of the most well known methods among actuaries. This can be considered as the first Bayesian approach to graduation, since it can be derived using Bayes theorem. But no real prior subjective information was formally used in the first development of this model, in contrast to the approach given by Klugman (1992) to the same model.

A model that could be seen as a step before the Whittaker one is the Kimeldorf and Jones (1967) model explained in London (1985). This model is written fully in

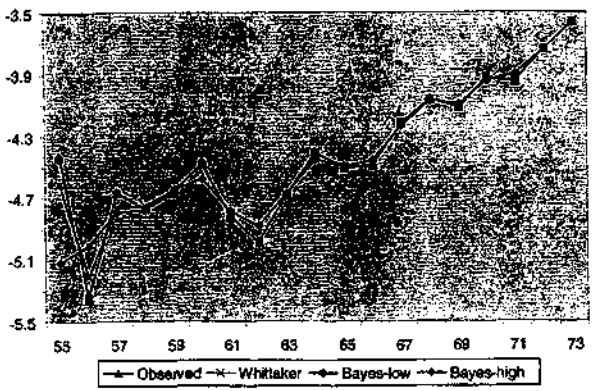
appendix C, and it states that the observed frequency of death will be modelled as:

$$X|\theta \sim \text{normal}\ (\theta, B)$$
$$\theta \sim \text{normal}\ (\mu, A) \qquad (14)$$

where $X = (x_1, \ldots, x_n)$,
$\theta = (\theta_1, \ldots, \theta_n)$,
$\mu = (\mu_1, \ldots, \mu_n)$,
n is the number of ages and A and B are known covariance matrices.

μ is taken from another life table and B is fixed and fully explained in the appendix C. The covariance matrix A is defined by the analyst and it is the one controlling the amount of smoothness. This is also presented in the appendix, but some other possible formats are discussed in London (1985). The graduated values are obtained as the posterior mean of θ and the graph of the estimates in the example analysed in London (1985) is shown on a log scale in figure 4.

Figure 4



The two Bayes results show how to control the model, with higher and lower level of smoothness, depending on the chosen value of A. Bayes-low is also so close to the observed data that it is even hard to distinguish them.

A different approach is presented in Klugman (1992), bringing a different approach to the Whittaker model. Instead of using prior information from another table, as in London (1985), a relationship is imposed among the parameters in θ. In order to do this, a design matrix is included transforming the model into:

$$X|\theta \sim \text{normal}\ (\theta, B)$$
$$K\theta \sim \text{normal}\ (0, A) \qquad (15)$$

Where K is the matrix that produces the z^{th} differences of a sequence of numbers. Choosing properly the values for A and B and letting $z = 3$ gives the posterior mean as the same solution as the one proposed by the Whittaker model. But in the new Whittaker approach not only the estimator of θ was found, but also its covariance matrix. In this way a confidence region could be easily found.

In fact, one of the first applications of the model expressed in formula (15) was the calculation of a reserve, where a confidence interval was also presented. The case when a prior distribution is given for A and B is also analysed in Klugman (1992).

Section 3 has given an outline of models with a Bayesian flavour. Models taking previous data information through $p(\theta)$ are shown, like Kimeldorf and Jones (1967) and the "pure Bayes with prior information" from Verrall (1990). Also models where $p(\theta)$ was presented to impose a dependency among θ was used, like the "empirical Bayes" from Verrall(1990) and Whittaker graduation from Klugman (1992). All the models have analytical solutions, and most of them are fully explained in the appendix.

Now, in the following section, an introduction to simulation will be given, which is the basis for the more elaborate models using Bayesian theory. Some models where the variance parameter also has a prior distribution will also be considered.

4 Simulation in Bayesian statistics

"The difficulty in carrying out the integration necessary to compute the posterior distribution has prevent such approaches from being seriously contemplated until recently (...)."
Carlin (1992)

As stated before, many models in Bayesian theory cannot be solved analytically. In order to apply them an approximation would be required and simulation is often used in implementing those models. This is not the only possible kind of approximation and the Gaussian quadrature used in Klugman (1992) could be cited as a different example. But since the advent of Gibbs sampling, simulation has superseded all other types of approximation.

In order to illustrate the simulation philosophy, suppose that the posterior of a specific parameter θ is needed. If an analytical solution were available, a formula would be derived, where the observed data and known parameters would be included, defining a final result. But, depending on the model, this solution will not be possible. In such cases an approximation for the posterior distribution of θ is needed. One way of finding this approximation is by simulation, that substitutes the posterior distribution by a large sample of θ based on the characteristics of the model. With this large sample of θ many summary statistics could be calculated, like the mean, variance or histogram, extracting from this sample of the posterior distribution all the

information needed.

There are a number of ways of simulating and in all of them some checking should be carried out to guarantee that the simulation set is really representative of the required distribution. For instance, it must be checked whether the simulation is mixing well or, in other words, if the simulation procedure is visiting all the possible values for θ. It should be also considered how large the sample should be, and whether the initial point where the simulation starts does not play a big role. Among many other issues, the moment when convergence to the true distribution of θ is achieved should also be monitored.

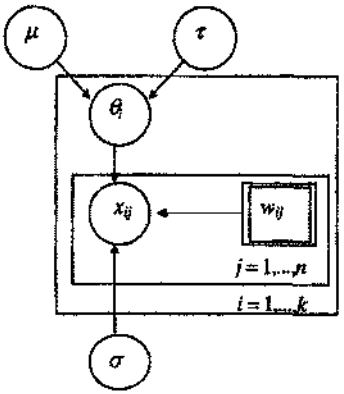
All these features can make the technique difficult to apply, and, even worse, perhaps dangerous to use. This happens because once all the needed procedures to start the simulation are ready, a sample of θ could always been obtained. This, however, does not mean that it is really representative of the posterior distribution. The only way the analyst could assure that the sample does not have any deviation from the posterior distribution is through the tests listed above.

The most popular type of simulation in Bayesian theory are the Markov chain Monte Carlo (MCMC) methods. This class of simulation has been used in a large number and wide range of applications, and has been found to be very powerful. The essence of the MCMC method is that by sampling from specific simple distributions (derived from the combination of the likelihood and prior distributions), a sample from the posterior distribution will be obtained in an asymptotic way. Among the techniques that use MCMC, one of the most popular is Gibbs sampling. WinBUGS or a specifically developed program could implement this method.

WinBUGS is the newest version of BUGS (Bayesian inference Using Gibbs Sampling) which was first made available in 1992. This software works under Microsoft Windows® and this makes it easier to manipulate. Many useful tools for analysis are already included, and this helps to check if the simulation follows the rules cited here before. There is also software called CODA that produces some tests to check whether the simulation can be regarded as representative of the posterior distribution. It also includes a manual and a set of examples, and the more interested reader should visit www.mrc-bsu.cam.ac.uk/bugs in order to get this free software. Although it has a very specific notation and use, someone interested in redoing the following examples should be able to do it, and get a feeling of what can be done under WinBUGS.

One way of representing the model which is the basis of WinBUGS is the graphical model. Such a scheme is often used in Bayesian analysis to give a better understanding of the models, particularly when the dependencies between the data and the parameters are complex. Figure 5 shows the graphical representation for the Bühlmann and Straub model described in subsection 3.1.

Figure 5



Where circles stand for random variables $(x_{ij}, \theta_i, \sigma^2, \mu, \tau^2)$, rectangles for constants (w_{ij}) and the big rectangles for the index (i and j). This graphical model shows that once the parameters θ_i are given, the data x_{ij} do not depend on μ or τ^2 any more. It also shows that once θ_i are given, they contain all the model information needed to update μ for instance. This feature is the basis of Gibbs sampling, since through this conditional independence it is possible to derive simple distributions, which will be used to update the parameters values. In the next sections most of the model will have a graphical model to guide the reader on applying the model in WinBUGS.

Simulation deals with missing values in a very straightforward way. Those values are treated as variables, in the same way as the parameters. So, in each iteration, a value for the missing value is also calculated and inference is carried out as usual. In WinBUGS, for instance, the missing value is stated as a "NA" (Not Available) in the data set itself.

In order to illustrate these techniques, the traditional models reviewed in section 3 will be reconsidered in this section. In all of them, simulation will be used. The order in which they are presented is also the same as in the previous section and some of the examples have their code in WinBUGS written in appendix D.

4.1 Credibility theory

Returning to the credibility model in subsection 3.1, two new models will be used here to apply WinBUGS. The first one only aims to show how to use WinBUGS and is defined by the simply addition of prior distributions for the unkoown parameters in the Bühlmann and Straub model. The second one changes the core

assumptions of the Bühlmann and Straub model and shows how this is easily implemented in a simulation environment.

The data are the same as the one in subsection 3.1 and were also analysed by Scollnik (1996) and Smith (1996). The approach used here is the one adopted in Pereira (1998).

Recalling the model from Bühlmann and Straub, $p(x_{ij}|\theta_i, \sigma)$ is normal$(\theta_i, \sigma^2/w_{ij})$ and $p(\theta_i|\mu, \tau)$ is a normal(μ, τ^2), with unknown σ^2, μ and τ^2. In the solution proposed by Klugman (1992) a set of non-informative prior distributions were used and the solution, which did not have a analytical solution, was found by a non-simulation technique. In that solution a program had to be specifically written in order to carry out the model implementation and, depending on the approximation technique chosen, the calculations could take 2 hours.

The first example in this subsection reanalyses those data, but using WinBUGS. The model is written in a WinBUGS terminology in figure 6 with the following set of prior distributions, which has a non-informative objective:

$$\begin{aligned} \mu &\sim \text{normal } (0, 10^5) \\ 1/\tau^2 &\sim \text{gamma } (0.001, 0.001) \\ 1/\sigma^2 &\sim \text{gamma } (0.001, 0.001) \end{aligned} \qquad (16)$$

Figure 6

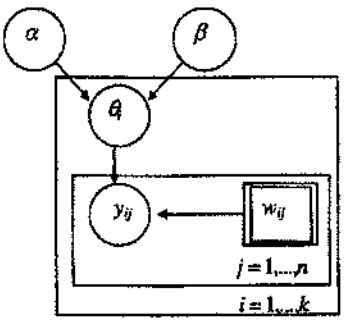
```
model BühlmannStraub; [12]
const
  N = 130,  # number of classes
  U = 6;   # number of observed years
var
  mu, theta[N],Y[N,U],tau,sigma,w[N,U],sigc[N,U];
  data in "datafile";
  inits in "initialfile";
{
  mu ~ dnorm (0, 1.0E-5);
  tau ~ dgamma (1.0E-3,1.0E-3);
  sigma ~ dgamma (1.0E-3,1.0E-3);
  for (i in 1:N) {
    theta[i] ~ dnorm (mi,tau);
    for (j in 1:U) {
      sigc[i,j] <-sigma*w[i,j];
      Y[i,j] ~ dnorm (theta[i],sigc[i,j]);
    }
  }
}
```

[12] In WinBUGS instead of the variance, the precision (1/variance) is used for the normal distribution

The implementation of this model took 5 minutes on a fairly old computer, with a total of 2500 simulations, where the first 500 were discarded to eliminate the effects of the initial conditions. Before showing the results the second model will be described. Since the observations are numbers of claims it is more suitable to model the data using, for instance, a Poisson distribution rather than normal distributions. In WinBUGS this is a direct generalisation of the previous model and it is only necessary to change the model, using non-informative prior distributions (also represented as graphical model in figure 7), to:

$$\begin{gathered} y_{ij} \mid \theta_i \sim \text{Poisson}(\theta_i \times w_{ij}) \\ \theta_i \mid \alpha, \beta \sim \text{gamma}(\alpha, \beta) \\ \alpha \sim \text{uniform}(0.01, 50) \text{ and } \beta \sim \text{uniform}(0.01, 50) \end{gathered} \tag{17}$$

Figure 7[13]



This model did not take much longer than the previous one to be implemented with the same amount of data. Since one of the main quantities of interest is the forecast of the number of claims for the 7th year, this is done in WinBUGS without calculating θ_i, but rather by sampling the value of y_{i7} directly. This is possible since the values for the 7th year can be treated as missing values. The table below gives the results, where the value of the deviance is related directly to the forecasted value of y_{i7}.

[13] WinBUGS code in appendix D.

Observed data			Normal		Poisson	
Class	w_{i7}	y_{i7}	Forecast	Deviance	Forecast	Deviance
4	-	0	-	-	-	-
11	229.83	8	10.15	7.28	10.21	3.57
112	18,809.67	45	38.24	66.06	35.64	6.63
70	54.81	0	0.60	3.47	0.254	0.57
20	1,315.37	22	41.24	16.84	41.48	6.55
89	79.63	40	29.56	4.11	32.82	6.17
Forecast error			13.22		12.44	

Comparing these values to the ones found in subsection 3.1, it is observed that the Normal solution is almost the same as the previous ones. The benefit for using the Poisson distribution can be seen in the smaller forecast error found in this case. And it is also observed that in many classes the deviance was smaller when the Poisson distribution was assumed.

4.2 Claims reserving

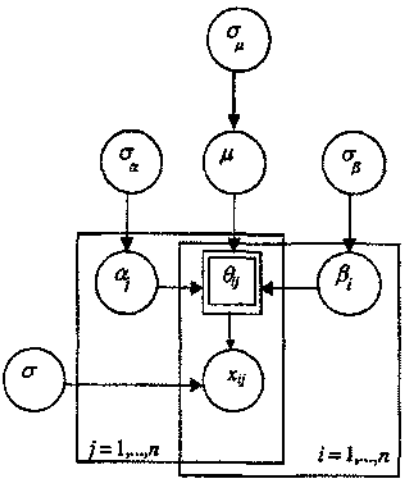
The flexibility in the solution by simulation gives an enormous number of models that can be applied to better understand processes in insurance. In the previous subsection the use of a Poisson distribution was a fairly easy and straightforward one.

The use of WinBUGS in order to implement the Gibbs sampling technique is a very convenient one. This is mainly because of the development of a specific program is not needed and the number of techniques to control the simulation which are already built in.

Not much research has been done in order to implement chain ladder based models using WinBUGS. This mainly due to the amount of missing values which there are in claims reserving (the outstanding claims are treated as missing values in WinBUGS). So in order to use such triangular data, the model was implemented either using specifically written programs, or by imposing very strong assumptions. Other researchers have used new models, which would not use the data in the triangular format, but the individual claim amounts. An overview of what has already been done in this direction will be given in subsection 5.3.

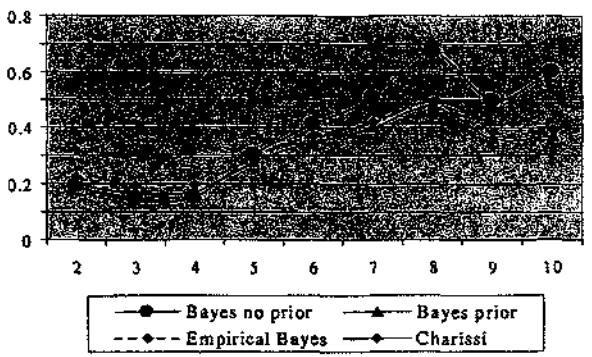
Two works using triangular data will be cited here. The first one is Charissi (1997) where the "pure Bayes without prior information" model in Verrall (1990) is reanalysed using BUGS (the previous version of WinBUGS). But now there is a proper prior distribution for each of the parameters. These are included in the second level, and independently of the chosen distribution, each one had to be centred on the values observed in the data, with quite a low variance. The graphical model would be as in figure 8:

Figure 8



The data from Taylor and Ashe (1983) were reanalysed and the results of the posterior mean for the row parameter is plotted in the figure 9 together with the values found before in Verrall (1990). On one hand, it is easy to see that the set of chosen prior was not able to influence much the mean of the row parameters (or even the other ones), keeping the same result as the one found in "pure Bayes with no prior". But, on the other hand, in this new analysis the influence of the prior was enough to decrease the standard error of the parameters by an average of 30% compared to the previous approach.

Figure 9



The second paper is Ntzoufras and Dellaportas (1997). Gibbs sampling is again used as simulation technique, but although this paper was prepared after the development of BUGS, a specific implementation program was used instead. Five models were presented in the paper and all of them were applied to the same data set. This set includes the inflation rate for the observed calendar years and two incremental development triangles: amount and number of claims. With all of this information in hand they proposed new models that would take into consideration the number of claims in order to predict the claim amounts, which would be deinflated before any analysis. Only one model among all five will be fully explained here. The more interested reader should report to the original paper in order to see all explanations and formulae for the other models.

"Log-normal & Poisson model" is a direct generalisation of Kremer (1982). Now, instead of using only the information from the amount of claims, the history of number of claim (n_{ij}) reported in row i and column j is also taken into consideration. Now the model in (11) will be changed to:

$$\begin{aligned} & x_{ij} = \ln(y_{ij}) \\ & x_{ij}|\theta_{ij}, \sigma^2 \sim \text{normal}(\theta_{ij}, \sigma^2) \\ & \theta_{ij} = \mu + \alpha_i + \beta_j + \ln(n_{ij}) \\ & n_{ij}|\lambda_{ij} \sim \text{Poisson}(\lambda_{ij}) \\ & \ln(\lambda_{ij}) = \mu^* + \alpha_i^* + \beta_j^* \end{aligned} \qquad (18)$$

with constraints and prior distributions for μ, α_i, β_j, μ^*, α_i^*, β_j^*, σ^2 fully described in the paper.

An analysis was performed with all models, and it was shown that for the specific data used the models that included also the number of claims, like the one explained above, had a better prediction than the ones that did not use such information. This was mainly due to the long tail characteristic of the data set, where claims were still being reported after 7 years of occurrence.

This subsection has shown that the flexibility of the simulation approach was able to allow also the inclusion of the development of number of claims in the chain ladder model. In the next subsection some applications of simulation will be used for graduation as well.

4.3 GRADUATION

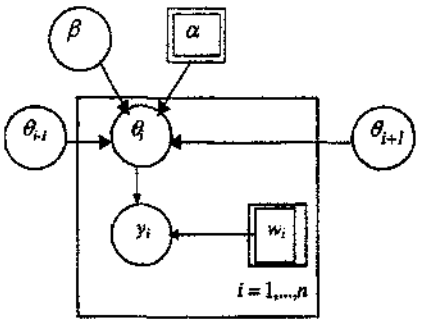
The paper from Carlin (1992) uses Gibbs sampling technique to graduate not only mortality table but also the aging factor cost related to health insurance. In both of these applications some restrictions were imposed in the model structure, like for instance, the growth on mortality expected in adulthood. Here only the mortality example will be explained.

The paper was developed before BUGS was implemented, so a specially written program carried out all calculations. In the graduation problem the data set has ages from 35 to 64, so 30 ages were observed. The model states that the number of deaths y_i in age i+34 for i=1,...,30 is Poisson distributed with intensity given by $\theta_i \times w_i$ where w_i is the number of policyholder in i. The model is written as:

$$\begin{aligned} y_i|\theta_i &\sim \text{Poisson}(\theta_i \times w_i) \\ \theta_i|\beta &\sim \text{gamma}(\alpha, \beta) \end{aligned} \qquad (19)$$

Where $\theta_i > 0$, $\theta_{30} < B$, $0 < \theta_2 - \theta_1 < \ldots < \theta_{30} - \theta_{29}$, B and α fixed, supposing a prior distribution for β. Now a graphical model is drawn for this model. It is shown in figure 10, where the imposed order among the parameters θ is also represented.

Figure 10



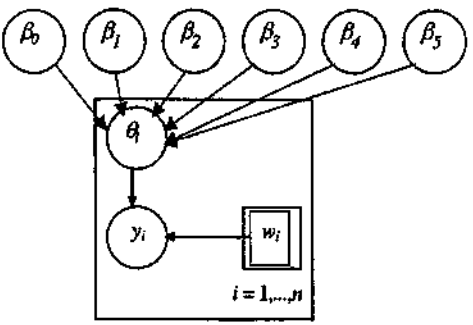
Some constraints were also imposed on the model and the more interested reader should refer to the original paper in order to see these in full. The results are also compared with the ones obtained by the Whittaker model and the author comments that "The Whittaker results are fairly similar to the Bayes results, though the Whittaker rates tend to be influenced more by the unusually low rate at age 63.". It means that the model was able to keep the growth among the parameters θ, although this was not observed for all ages in the data set.

An application of BUGS to graduation can be found in Kouyoumoutzis (1998). In this work a number of models were investigated and the one explained here is based on a third degree polynomial regression analysis and expressed by:

$$\begin{gathered} y_i|\theta_i \sim \text{Poisson}(w_i \times \theta_i) \text{ with} \\ \ln(\theta_i) = \beta_0 + \beta_1 i + \beta_2(3i^2-1)/2 + \beta_3(5i^3-3i)/2 + \\ \beta_4(35i^4-30i^2+3)/8 + \beta_5(315i^5-350i^3+75i^2)/40, \\ \text{and } \beta_j \sim \text{normal}(0,0.001) \text{ for } j=0,\ldots,5 \end{gathered} \qquad (20)$$

The time needed to run the simulation was again very small and the smoothed values fitted well the data. The graphical model is show below in figure 11.

Figure11[14]



In this section a review of traditional models revised in a Gibbs sampling approach has been given. Different, new models were incorporated by the inclusion of simulation into the modelling process. It is expected that the more actuaries are able to use WinBUGS, and more generally Gibbs sampling, the more revisions of traditional models will emerge.

In the next section completely new ideas will be presented. The assumptions used in macro models are completely dropped and models with approaches closer to the process itself will be used.

5 New areas, new possibilities

" (...) the potential for these methods in insurance application is great."
Boskov and Verrall (1994)

Up to this point we have discussed well known models which were rewritten in order to give a Bayesian approach. This opens a broad area of research to Bayesian theory, since most of the well established models in actuarial science can be reviewed in a Bayesian way. And with the advent of simulation, solutions can be found for most of them.

But one of most appealing features of a Bayesian analysis is the broader set of models that can be built, models which do not have a classical equivalent approach. This feature is mainly due to the simulation advanced on the last few years, when some new models were developed. In this section some of those new models are described, including some practical appealing ideas which are easily embraced by a

[14] WinBUGS code in appendix D.

Bayesian model.

It may turn out (and this is something that remains to be seen) the most important of these new ideas is the ability to model at the individual policy level. Now an "engineering approach", when assumptions are made straight in the process itself rather than on the aggregated data, fits fairly easily within a Bayesian model.

In order to show how it is done, three examples are presented in the next subsections. The first is the use of spatial models in the rating by area problem, not using individual data but only the loss ratio and exposure by area. The individual data will be considered in the second model, which is an aggregation of continuous variables in the problem of transforming ages into factors in the rating process. And the last example is an application to claims reserving, but now considering the individual data, instead of the usual triangular format.

All three models use the simulation approach, but none could use WinBUGS and a specific implementation program had to be written. Their formulae will not be described in detail, but their assumptions are fully explained. It is hoped that the reader could get a feeling of those models here, and should refer to the original papers for a full formulation.

5.1 Rating by postcode area

There are many factors that could influence the frequency or cost of a claim and that should be taken into consideration when defining the value of the premium. One of these is the area where, for example, a car is used or parked most often and this characteristic is usually taken into account through the neighbourhood where the policyholder lives.

Neighbourhood could have many interpretations, but here postcode is used. In an office environment it is common to aggregate postcodes with similar experiences in the same class. At the end of this procedure a small number of classes will be derived, but the vicinity information is not formally taken into account by the model.

Taylor (1989) published the first paper with some statistical basis, which addressed how to carry out this aggregation using the vicinity information. He adapted a two-dimension splines model to the postcode problem, with a totally non-Bayesian approach.

In this paper a review of Boskov and Verrall (1994) will be presented. They use a Bayesian approach, applying spatial models mainly used in epidemiology and satellite image restoration among other fields. The basis for such models is that areas that are close together are more likely to be similar in risk than areas that are far apart.

The aim of the model is to find a value for risk parameter (θ_i), that will be smoothed oven the whole area (that contains n postcodes) but considering only

information from its neighbours. The data contain the observed loss ratio (x_i) for each postcode area i, and they are assumed to have a normal distribution as follows:

$$x_i | \theta_i \sim \text{normal}(\theta_i, \sigma^2 w_i) \qquad i = 1,\ldots,n \qquad (21)$$

Where w_i is the exposure for postcode area i. Instead of using the variance as a variable like is some models seen before is this paper, σ will be a constant chosen by the analyst fixing the required level of smoothness. The bigger σ, the smoother is the result for the posterior mean of θ_i.

The most important idea of the model comes in the definition of the second level of distributions, when a relationship among the risk parameters θ_i is defined. For each postcode risk θ_i an adjacency set is defined as in figure 12, where the darker areas are included in the neighbourhood of the risk.

Figure 12

So the risk parameter of each postcode is defined to be normally distributed, centred on the average of all risk parameters in the adjacency set. All risk parameters are defined at the same time, influencing their neighbours as well.

This model does not have a possible analytical solution, and a simulation approach was used in order to find the posterior of θ_i. A MCMC method was used and the full model explanation can be found in the original paper. In there an analysis of the results are shown for different levels of smoothness, and it is really interesting to observe that the model did work. The risk parameters really took some information from the neighbours.

In the following subsections, models considering individual data will be presented. Their solutions are derived through simulation, and one of their common feature is the long time needed to perform the implementation. This could be a barrier to a practical use, but their benefits could easily justify the time spent on them.

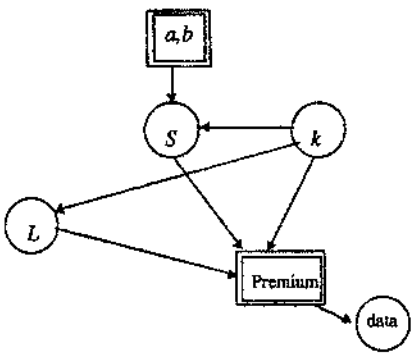
5.2 GROUPING AGES

In Pereira and Verrall (1999) a method of transforming a continuous variable into fewer factors is presented. The paper works in the particular example of policyholder age, but the model could be applied in any other kind of continuous variable.

The main objective of transforming age into a factor is to summarise information. Actuaries usually do this when, for instance, age is used as a covariate in ratemaking. Usually age is considered as an integer number (which is already a year aggregation) and the transformation to groups is a further process. The first step, considering age in years, is done without a proper analysis and could bring distortions to the group definition. In this new approach the pure data are used, dropping the first step used before, and transforming the real age into a factor with only a few classes.

Informally, the model is specified in the following way. Suppose that age is limited to some interval [a,b]. The procedure would find at the same time how many intervals (k) there should be in [a,b], where they would be best located ($S=(s_1,\ldots,s_{k-1})$) and what risk intensity ($L=(l_0,\ldots,l_{k-1})$) are appropriate for each of them. This approach is based on the premium philosophy that once we know the groups, the premium level will be the same for any policyholder included in a particular group. The graphical model is presented in figure 13.

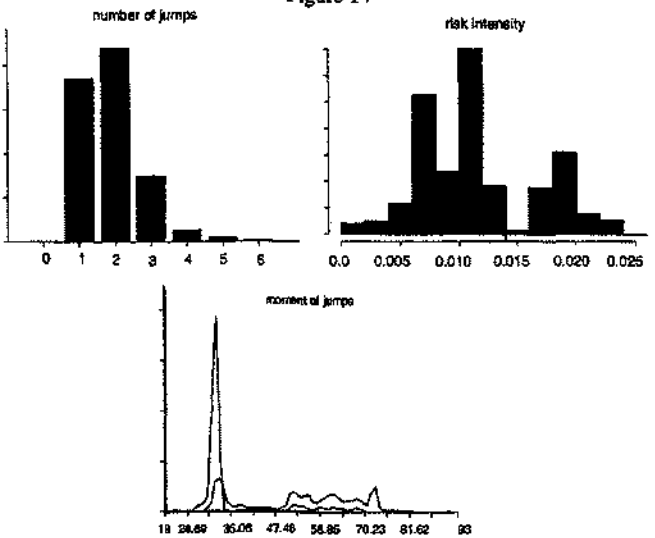
Figure 13



Finding k, S and L at the same time, would mean that the size of S and L changes according to the value of k. This makes the model fairly difficult to be applied and a generalisation of the MCMC techniques is used. It is called Reversible Jump Markov chain Monte Carlo (RJMCMC) and in this procedure the size of k can be changed in each iteration.

Pereira and Verrall (1999) uses as a case study data from a bodily injury motor insurance. The data consist of the number of claims on an individual basis, their date of occurrence and the age (in days) of the policyholder, which was some number in the interval (in years) [19.39, 93.89]. The solution for the model was found through the calculation of the posterior distribution based on the sample obtained. The plot of the posterior distributions for k, L and S are presented in figure 14.

Figure 14



With these posterior distributions at hand, the analyst should choose the values for the parameters. For the number of jumps, k=2 is the mode of the posterior distribution and can be seen as a proper solution. Since the other distributions have more than one mode (as expected), such an approach is not as clear as in the definition of k and many analyses could be used. In the original paper the posterior mode was calculated.

In this example the step of defining an estimator after finding the posterior distribution has been considered. This could be good or bad. On one hand, the analyst does not have a closed and final solution, and he is able to draw conclusions based on a distribution, which gives an enormous amount of information. But on the other hand, different analysts could chose different values, based on the same result.

Now an example with more specific answers will be presented. This will be the last model explained in this section, giving an interesting approach to how include

individual information in the claims reserving procedure.

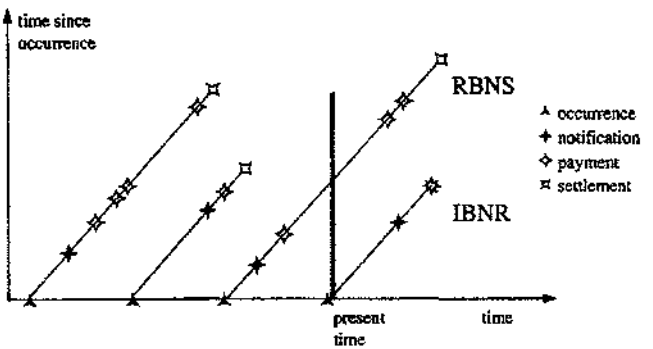
5.3 CLAIMS RESERVING

The majority of methods for estimating claims reserves are based on macro models, where the data are aggregated in a triangular format like the chain ladder model. Micro models, where the individual policyholder characteristics are statistically taken into consideration are not usual. The only office based procedure that takes into consideration some individual information is the case reserve definition, when the claim characteristic is used, but this does not have any statistical basis.

One of reasons why this individual characteristic is not used in statistical models could be the difficulties that surround any calculation on an individual claim basis. The fluctuation related to any individual estimate, could be also a good justification for the lack of use of such models. The key question would be to use such information, but in a more robust way.

Before analysing the model, it would be helpful to think of the claim process. Consider now the analysis done by Norberg (1993) and represented in figure 15. Such a scheme shows how each claim could be different from another. Most claims can not be completely settled at the moment the claims reserve is calculated (Reported But Not Settled - RBNS) and after a claim is incurred but not reported (IBNR). It also highlights the partial payment process, which for most type of insurance is more usual than a simple payment.

Figure 15



Using this approach Haastrup and Arjas (1996) proposed a new method, using a Bayesian analysis to define claims reserve for the whole portfolio, but considering individual information. The IBNR and RBNS claims reserves are calculated

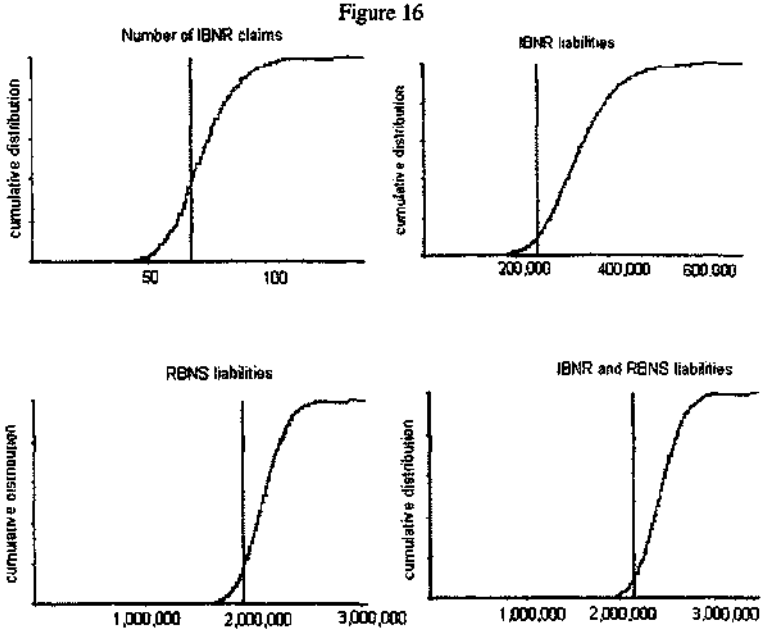
separately. In this model no information from the claim itself was taken into consideration, but it is possible to do so in such a framework.

In their model the claims frequency and severity are modelled separately. In the first model age, sex, report delay and calendar time of occurrence are included, and in the second model the analysis uses partial payments. MCMC simulation is used in order to obtain the estimated posterior distributions.

The way of handling missing values in a Bayesian framework is also explored and the IBNR claims are considered as missing. Since simulation is used, it is possible to sample at each step the number of claims that had already occurred and that are missing (IBNR) and their correspondent amounts. At the end of the simulation a sample of IBNR numbers and values is available and its posterior distribution can be approximated. The amount of RBNS claims is calculated in the same way.

The result of this model is given in figure 16, where the graphs are produced for the number and amount of IBNR, amount of RBNS and both liabilities together. Values are shown in Danish currency.

Figure 16



This model suggests many ideas for further development. If individual

information could be taken into account in a statistical model, it means that characteristics of the claim itself could also be formally considered.

And since their approach also considers the calendar time in order to define the reserves, it is possible to obtain the amount of the reserve at a specific moment in time. The vertical line included in the graph shows this feature, giving the correspondent values expected in a specific evaluation moment.

6 Conclusions

" (...)actuaries having spent the past half-century seeking linear solutions (...) logical solution is to drop the linear approximation and seek the true Bayesian solution."
Klugman (1992)

As it was shown throughout this paper, many models have been developed in a Bayesian framework. Some of them were only an extension of well known models, but others included new ideas to the actuarial analysis. And we have given only a sample of a large variety of papers. It is hoped that this paper has excited the curiosity of actuaries and that Bayesian theory will be also applied in practice.

Bayesian models bring to the practitioner actuary two attractive possibilities: formal inclusion of judgement and wider number of models. In many cases, these characteristics are very interesting for any actuary. But how can they be made really practical? In order to apply them the following steps should be done. Firstly, it is important to make sure that Bayesian theory is fully understood. Secondly, simulation by MCMC should also be covered.

After these two main areas are covered anyone could easily apply the models explained in section 3 and 4 of the present paper. The reader is also encouraged to experiment with WinBUGS, which is a powerful tool in Bayesian analysis. WinBUGS could solve even methods with analytical solution, since it is fast and accurate.

The range of applications in actuarial science where Bayesian theory could be used is enormous. For instance, a prior distribution for the interest rate could be included in a pension fund analysis, extreme value theory[15] could be used to price a catastrophe bond and information on the claim itself could be included in the model described in subsection 5.3 for reserving. An application that seems straightforward is the inclusion of benchmarks in the chain ladder model described in subsection 3.2 and 4.2.

[15] Definition is in the reference list on Bayesian Theory at the end of this paper.

References:

_____ (1998) WinBUGS Manual; MRC Biostatistcs Unit Cambridge.

_____ (1998) WinBUGS Examples; MRC Biostatistcs Unit Cambridge.

Arjas, E. and Haastrup, S. (1996) Claims reserving in continuous time; a nonparametric Bayesian Approach. ASTIN Bulletin, vol 26, no.2, pp 139-164.

Bailey, A (1950) Credibility procedures, Laplace's generalisation of Bayes' rule and the combination of collateral knowledge with observed data; Proceedings of the Casualty Actuarial Society, vol 37.

Boskov, M. and Verrall, R. J. (1994) Premium rating by geographic area using spatial models; ASTIN Bulletin, vol 24, no. 1, pp 134-143.

Bühlmann, H and Jewell (1987) Hierarchical credibility revisited; Bulletin of the Association of Swiss Actuaries.

Bühlmann, H and Straub (1970) Credibility for loss ratios; Bulletin of the Association of Swiss Actuaries, vol. 70.

Bühlmann, H. (1967) Experience rating and credibility; Astin Bulletin, vol 4.

Carlin, B.P. (1992) A simple Monte Carlo approach to Bayesian graduation; Transactions of Society of Actuaries, vol XLIV, pp 55-76.

Chrissi, D. (1997) Claims reserving under a Bayesian approach using BUGS; Master dissertation; City University: London.

Dellaportas, P. and Ntzoufras, I. (1997) Bayesian prediction of outstanding claims. University Report.

DeGroot, M.H. (1986) Probability and statistics: second edition; Addison-Wesley publishing company: USA.

Haberman, S. (1996) Landmarks in the history of actuarial science (up to 1919); Research report, City University: London/UK.

Kimeldorf, G.S. and Jones, D.A. (1967) Bayesian graduation; Transactions of Society of Actuaries, vol XIX, pp 66.

Klugman, S. A. (1992) Bayesian statistics in actuarial science with emphasis on credibility theory; Boston: Kluwer.

Kouyoumoutzis, K (1998) Monitoring mortality over time; Master dissertation; City University: London.

Kremer (1982) Exponential smoothing and credibility theory; Insurance: Mathematics and Economics, vol 1, n° 3.

Liu,Y.-H, Makov, U. E. and Smith, A. F. M. (1996) Bayesian methods in actuarial science; The Statistician, 45, n° 4, pp 503-515.

London, D. (1985) Graduation: the revision of estimates; ACTEX Publications: USA.

Norberg, R (1993) Prediction of outstanding liabilities in non-life insurance. Astin Bulletin, vol 23, n° 1, 95-115.

Pereira, F.C. (1998) Teoria da credibilidade: uma abordagem integrada. Caderno tese. Funenseg: Rio de Janeiro/BR.

Pereira, F.C. and Verrall, R. J. (1999) A Markov chain Monte Carlo approach to grouping premium rating factors. (to appear).

Scollnik, D.P.M. (1996) An introduction to Markov chain Monte Carlo methods and their actuarial applications. Proceedings of the Casualty Actuarial Society, vol

LXXXIII, n° 158.
Taylor, G.C. (1989) Use of spline functions for premium rating by geographic area. Astin Bulletin, vol 19, n° 1, pp 91-122.
Taylor, G.C. and Ashe, F.R. (1983) Second moments of estimates of outstanding claims. Journal of Econometrics, vol 23, pp 37-61.
Verrall, R. J. (1990) Bayes and empirical Bayes estimation for the chain ladder model. Astin Bulletin, vol 20, n° 2, 217-243.

References on Bayesian Theory:

Chib, S. and Greenberg, E. (1994) Understanding the Metropolis-Hastings algorithm. University Report.
Gamerman, D and Migon, H. (1993) Inferência estatística: uma abordagem integrada; Mathematics Institute, Federal University of Rio de Janeiro - UFRJ.
Gamerman, D. (1997) Markov chain Monte Carlo: stochastic simulation for Bayesian inference. London: Chapman & Hall.
Gilks, W.R., Richardson, S. and Spiegelhalter,D. J. (1996) Practical Markov chain Monte Carlo. London: Chapman & Hall.
Green, P. J. (1995) Reversible jump MCMC computation and Bayesian model determination. Biometrika, 82, 711-732.
Green, P. J. and Richardson, S. (1997) On Bayesian analysis of mixtures with an unknown number of components. J.R.Statistic Society B, 59, n° 4, pp 000-000.
Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika, vol 57, pp 97-109.

References on Credibility Theory:

_____ (1986) Special issue on credibility theory; Insurance Abstract and Reviews vol. 2, Feb.1986, n° 3.
Albrecht (1985)An evolutionary credibility model for claim numbers; Astin Bulletin, vol. 15, n°1.
Dannenburg, D. (1993) Some results on the estimation of the credibility factor in the classical Bühlmann model; XXIV Astin Colloquium.
Goovaerts, M.J and Hoogstad, W.J. (1987) Credibility theory; Surveys of Actuarial Studies, n° 4, Nationale-Nederlanden N.V.
Goovaerts, M.J, Kaas, R., Van Heerwaarden, A.E. and Bauwelinckx, T. (1990) Effective actuarial methods; Elsevier Science Publishing Company, Holland.
Hachemeister (1975) Credibility for regression models with application to trend; Credibility: theory and applications, Proceedings of the Berkeley Actuarial Research Conference on credibility, Academic Press.
Jewell (1974) Credible means are exact Bayesian for exponential families; Astin Bulletin,vol.8,n°1.
Jewell (1975) The use of collateral data in credibility theory: a hierarchical model; Giornale dell'Istituto Italiano degli Atuari, vol 38.
Jewell (1976) A survey of credibility theory; Operations Research Center, Research

Report n° 76 - 3, Berkeley.
Jong and Zehnwirth (1983) Credibility theory and the Kalman filter; Insurance: Mathematics and Economics, vol 2.
Kling, B (1993) A note on iterative non-linear regression in credibility; XXIV Astin Colloquium.
Ledolter, J. , Klugman, S. and Lee, C.S. (1990) Credibility models with time-varing trend components, Astin Bulletin, vol 21, n° 1.
Longley-Cook (1962) An introduction to credibility theory; Proceedings of the Casualty Actuarial Society, vol 49.
Sundt (1982) Invariantly recursive credibility estimation; Insurance: Mathematics and Economics, vol 1, n° 3.
Sundt (1983) Finite credibility formulae in evolutionary models; Scandinavian Actuarial Journal, n°2
Sundt (1987) Credibility estimators with geometric weights; XX Astin Colloquium Scheveningen.
Waters, H. R. (1987) Special note: an introduction to credibility theory; Institute of Actuaries and Faculty of Actuaries.
Whitney, A. (1918) The theory of experience rating; Proceedings of the Casualty Actuarial Society, vol 4.

Appendix A

General model expression

$X\|\theta = A\theta + \varepsilon_1$	$\varepsilon_1 \sim \text{normal}(0, F)$
$\theta = B\mu + \varepsilon_2$	$\varepsilon_2 \sim \text{normal}(0, G)$
$p(\mu, F, G)$	respective priors

	BÜHLMANN(1967)	BÜHLMANN & STRAUB(1970)
Expression	$x_{ij} = \theta_i + \varepsilon_{ij}$ $\varepsilon_{ij} \sim \text{normal}(0, \sigma^2)$ $\theta_i = \mu + \varepsilon_i$ $\varepsilon_i \sim \text{normal}(0, \tau^2)$ $p(\mu, \sigma^2, \tau^2)$	$x_{ij} = \theta_i + \varepsilon_{ij}$ $\varepsilon_{ij} \sim \text{normal}(0, \sigma^2/w_{ij})$ $\theta_i = \mu + \varepsilon_i$ $\varepsilon_i \sim \text{normal}(0, \tau^2)$ $p(\mu, \sigma^2, \tau^2)$ w_{ij} is the weight
X'	$(x_{11},\ldots,x_{1n},x_{21},\ldots,x_{kn})$	$(x_{11},\ldots,x_{1n},x_{21},\ldots,x_{kn})$
θ'	$(\theta_1,\ldots,\theta_k)$	$(\theta_1,\ldots,\theta_k)$
B	1_k	1_k
F	$\sigma^2 I_{(kn)}$	$\text{diag}(\sigma^2/w_{11},\ldots,\sigma^2/w_{kn})$
μ	$\mu 1_k$	$\mu 1_k$
A	$\begin{bmatrix} 1_n & 0_n & \cdots & 0_n \\ 0_n & 1_n & \cdots & 0_n \\ \vdots & \vdots & \ddots & \vdots \\ 0_n & 0_n & \cdots & 1_n \end{bmatrix}$	$\begin{bmatrix} 1_n & 0_n & \cdots & 0_n \\ 0_n & 1_n & \cdots & 0_n \\ \vdots & \vdots & \ddots & \vdots \\ 0_n & 0_n & \cdots & 1_n \end{bmatrix}$
G	$\tau^2 I_k$	$\tau^2 I_k$

Usual estimators for variances used in the classical approach:

σ^2	$\frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$	$\frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} (x_{ij} - x_{iw})^2$ $x_{iw} = \sum_{j=1}^{n} \frac{w_{ij}}{w_{i\cdot}} x_{ij}$
τ^2	$\frac{1}{k-1} \sum_{i=1}^{k} (\bar{x} - \bar{x}_i)^2 - \frac{\hat{\sigma}^2}{n}$	$\frac{w_{\cdot\cdot}}{w_{\cdot\cdot}^2 - \sum_{i=1}^{k} w_{i\cdot}^2} \left(\sum_{i=1}^{k} w_{i\cdot} (x_{iw} - x_{ww})^2 - (k-1)\hat{\sigma}^2 \right)$ $x_{ww} = \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{w_{ij}}{w_{i\cdot}} x_{ij}$

Appendix B

Chain Ladder models (subsection 3.2)

General formula for the models presented at Verrall(1990):

$$X \mid \theta \sim \text{normal}(K\theta,\ \sigma^2 I);$$
$$\theta \mid \theta_I \sim \text{normal}(\theta_I, \Sigma);$$

with

X, θ, K, I, Σ and σ^2 as described in equation (12) and

	Pure Bayes without prior information	Pure Bayes with prior information
Assumption	$\sigma_\mu^{-2} = \sigma_\alpha^{-2} = \sigma_\beta^{-2} = 0$	$\theta_I = (0,0.3,\ldots, 0.3,0,\ldots, 0)$; $\sigma_\mu^{-2} = \sigma_\beta^{-2} = 0$; $\sigma_\alpha^{-2} = 0.05$
Posterior $p(\theta \mid X)$	normal$(\hat{\theta}, [\sigma^{-2}K'K]^{-1})$	normal$(z\hat{\theta} + (I-z)\theta_I, [\sigma^{-2}K'K + A^{-1}]^{-1})$
Formulae	$K\hat{\theta} = X$ $\hat{\sigma}^{-2} = \frac{(X - K\hat{\theta})'(X - K\hat{\theta})}{(n+2)}$	$\hat{\theta}$ and $\hat{\sigma}^{-2}$ as in model 1 $z = [\sigma^{-2}K'K + A^{-1}]^{-1}[\sigma^{-2}K'K]$ $A^{-1} = \begin{bmatrix} 0 & & & & \\ & 500 & & & \\ & & \ddots & & \\ & & & 500 & \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix}$

Appendix C

Kimeldorf and Jones (1967) model (formulae taken from London(1985))

posterior distribution:

$\theta \mid X \sim \text{normal}((A^{-1}+B^{-1})^{-1}(B^{-1}X + A^{-1}\mu), (A^{-1}+B^{-1})^{-1})$

$\mu = (m_1, m_2, \ldots, m_n)$

$B = \text{diag}(b_{11}, \ldots, b_{nn})$

$b_{ii} = \frac{m_i(1-m_i)}{n_i}$, where n_i is the number of policyholders in age i.

A has elements $a_{ij} = p^2 r^{|j-i|}$ (with r and p defined by the analyst)

Appendix D

WinBUGS code for figure 7

```
model poisson;
const
  N = 130,  # number of classes
  Q = 895;  # number of observation
var
  mu, tau,theta[N],Y[Q],P[Q],the_temp[Q],T[N],I[N];
  data in  "d27nna.txt";
  inits in "d27nna.in";
{
  mu ~ dunif(0.01,50);
  tau  ~ dunif(0.01,50);
  for (i in 1:N) {
    theta[i] ~ dgamma(mi,tau);
    for (j in I[i]:T[i]) {
      theta[j] <-the_temp[i]*w[j];
      Y[j] ~ dpois(theta[j]);
    }
  }
}
```

WinBUGS code for figure 11

```
model graduation;
const
  N = 68;  # number of ages
var
  W[N], Y[N], age[N], theta[N], theta_temp[N],
  b0, b1, b2, b3, b4, b5;
  data in  "datafile";
  inits in "initialfile";
{
 for (i in 1:N) {
    Y[i] ~dpois(theta_temp[i]);
    theta_temp[i] <- W[i]*theta[i];
    ln (theta[i]) <- b0+b1*age[i]+
    b2*((3*pow(age[i],2)-1)/2)+
    b3*((5*pow(age[i],3)-3*age[i])/2)+
    b4*((35*pow(age[i],4)-30*pow(age[i],2)+3)/8)+
    b5*((315*pow(age[i],5)-350*(age[i],3)+75*pow(age[i],2))/40);
}
b0 ~ dnorm (0, 0.001);
b1 ~ dnorm (0, 0.001);
b2 ~ dnorm (0, 0.001);
b3 ~ dnorm (0, 0.001);
b4 ~ dnorm (0, 0.001);
b5 ~ dnorm (0, 0.001);
}
```