BAYESIAN ANALYSIS OF AN AGGREGATE CLAIM MODEL USING VARIOUS LOSS DISTRIBUTIONS

by
Claire Dudley

A dissertation submitted for the award of the degree of
Master of Science in Actuarial Management

School of Mathematical and Computer Sciences
Actuarial Mathematics & Statistics
Heriot-Watt University Edinburgh

ABSTRACT

In the general insurance market, insurers need to use the data gathered from previous years of experience to make predictions about future liabilities. The statistical analysis can be performed using Bayesian methodology, which is increasingly common in actuarial science. We apply Bayesian methods to an aggregate risk model using simulation based procedures for estimation and use predictive distributions to make inferences about future aggregate claims. We examine and compare different forms of the loss distribution that may be used in the aggregate risk model, and investigate issues surrounding the application of Bayesian analysis and Monte Carlo Markov Chain methodology in this context.

BIBLIOGRAPHY

[1] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.

[2] Casella, G. & Berger, R.L. (1990). *Statistical Inference*. Wadsworth and Brooks/Cole, Pacific Grove, California.

[3] Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. *Biometrika*, **45**, 293-315. (Reprinted from *Phil. Trans. Roy. Soc.*, **53** (1763), 370-418, with biographical note by G.A. Barnard.)

[4] Laplace, P.S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

[5] Jeffreys, H. (1939). *Theory of Probability*. OUP.

[6] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. A*, **231**, 289-337.

[7] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machine. *J. Chem. Phys.*, **21**, 1087-1091.

[8] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, **6**, 721-741.

[9] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds. (1996*). Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

[10] Makov, U.E., Smith, A.F.M. & Liu, Y.-H. (1996). Bayesian methods in actuarial science. *The Statistician* **45** (4), 503-515.

[11] Rytgaard, M. (1990). Estimation in the Pareto distribution. *ASTIN Bulletin* **20** (2), 201-216.

[12] Panjer, H.H. & Willmot, G.E. (1983). Compound Poisson Models in Actuarial Risk Theory. *Journal of Econometrics* **23**, 63-76.

[13] Hogg, R.V. & Klugman, S.A. (1984). *Loss Distributions*. Wiley, Toronto.

[14] http://www.mrc-bsu.cam.ac.uk/bugs

[15] Pai, J.S. (1997). Bayesian Analysis of Compound Loss Distributions. *Journal of Econometrics* **79**, 129-146.

[16] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-472.

[17] http://www.mrc-bsu.cam.ac.uk/bugs/classic/coda04/readme.shtml

[18] Barnett, V. (1973). *Comparative Statistical Inference*. Wiley.

[19] Gibson, G.J. Notes on Bayesian Statistics. Heriot-Watt University.

BAYESIAN ANALYSIS OF AN AGGREGATE CLAIM MODEL USING VARIOUS LOSS DISTRIBUTIONS

SUMMARY

In this dissertation, I investigated ways of modelling aggregate claims in general insurance. Insurers need to know the expected level of claims they face in future years, and percentiles of the aggregate claim amount so that they can estimate the level of reserves required to be held in respect of a certain class of business. One way of doing this is by using an aggregate claim model. This aims to fit a probability distribution to the number of claims received in a year and another distribution to the size of the individual claims received and combine them in such a way that an approximate distribution can be found to model the aggregate amount of claims. We can then use this aggregate distribution to estimate percentiles and hence the level of solvency reserves required in future years.

In order to fit distributions to the data, I used Bayesian methodology. This approach assumes that all parameters in the distribution are themselves variables. Bayesian methods are very useful in actuarial science as they enable us to learn about the whole distributions of quantities rather than just obtain an expected value for each parameter. They allow us to include many levels of randomness in the analysis through the use of prior distributions for each parameter, which highlights the uncertainty regarding individual distributions or parameters and enables us to incorporate any prior knowledge we have. In addition, we can update the posterior distribution obtained when new information becomes available, which helps us to increase the reliability of our estimates. The posterior distribution can be used to give us an indication of the expected value of the parameter and the reliability of the inference (by considering the variance of the distribution). With the advent of computing technology, Bayesian methods are becoming ever easier to use, which has resulted in Bayesian methods becoming more prevalent than ever before.

To apply the Bayesian methodology, I used simulation-based techniques from the WinBUGS software. This implements the Gibbs sampling technique (a special case of the Metropolis-Hastings algorithm for Monte Carlo Markov Chains) in the case where

we use conjugate priors, to obtain Bayesian inference. To use the software, we need to specify the model to be used, the data to analyse and some initial values for the parameters and then wait for equilibrium to be reached. It then outputs distributions for each parameter, which can be used to give the posterior distributions of the quantities under investigation. As part of this dissertation, I explained several different issues surrounding testing the convergence of the Markov chains used in simulation. To use the posterior distributions output by the program with confidence, we need to be sure that equilibrium has been attained. The WinBUGS software provides several ways to test the extent to which this has happened.

The data I analysed is provided in 'Estimation in the Pareto distribution' by Rytgaard and relates to the loss amounts from a portfolio of motor insurance policies with an excess. The methods used in this project could, however, be applied to any similar set of claims data. I used a Compound Poisson model for the aggregate claim (although other discrete distributions could be used to model the number of claims) and a variety of loss distributions to model the size of individual claims. I compared the results derived by using Bayesian methodology with those from classical statistics to see which had the best fit with the data.

The loss distributions that I investigated included both standard distributions (such as the Exponential, Gamma, Pareto, Lognormal and Weibull) and their truncated forms (as the data was from a form of truncated distribution due to the excess being applied). My findings showed that for these data, using a vague Gamma prior, the Pareto distribution had the best fit. This result was not surprising as the Pareto model is often used to model individual claim sizes, particularly where an excess is involved. The prior distributions I used were vague, as I had no information regarding the actual distribution of the parameters, and of Gamma form, as that is conjugate to the Poisson-Pareto model. In general, among the other loss distributions, those that were truncated had a better fit than the standard forms. There are many other, more general, loss distributions that could also be tested, including those with a larger number of parameters, and one of these could have an even better fit to the data.

I then considered the sensitivity of my results to the form of the prior distribution used. For this, I only used the Poisson-Pareto model (as it had the best fit of the models I had examined). The priors I used were all vague (to reflect the lack of

knowledge about the parameters), but of different forms. Since the results showed similar expected values for the parameters, I concluded that the Pareto model is relatively robust to the form of prior used.

Next, I compared the distributions for individual claim size obtained from Bayesian analysis with those found by using the Maximum Likelihood Estimator and Method of Moments Estimator from classical statistics. In general, the Bayesian distributions had a much closer fit to the data provided than the classical ones. Since the Bayesian methods had provided us with full distributions for the variables, rather than a point estimator, we could then use predictive distributions to find a model for the aggregate claims.

Using the Poisson-Pareto model as calculated with the Bayesian techniques, I then simulated values from the aggregate distribution using two different methods, to help us to obtain a numerical distribution for the aggregate claims. The simulations both involved simulating a number of claims for the year, and then simulating the size of each claim before summing them to achieve the aggregate. I found that the shape of the aggregate claims was broadly similar in shape to a loss distribution, but that none of the loss distributions I tried had a particularly close fit, due to certain features of the aggregate claims distribution caused by the existence of the excess in the data. If needed, we could fit an analytical formula to the aggregate claims distribution (or even just to the tail), which would help in estimating points in the tail, and hence setting the amount of reserve that the insurer would need to hold.

TABLE OF CONTENTS

INTRODUCTION

Insurance is, by nature, a very uncertain subject. Insured events occur at random times and, particularly in the field of general insurance, the amounts of the claims are also random. The act of taking out insurance relieves the insured parties of some of the risk involved, instead passing it on to insurers, in return for a stable series of payments – the premium. The insurer must calculate the value of premium it should charge, which will be related to the total expenditure it is likely to have in fulfilling the conditions of the policies. In addition to this, the insurer must ensure it has sufficient funds, or reserves, in place to pay out claims when they arrive. In order to do this, they need to learn about not only the average amount to be paid out in any one year (which would be sufficient to determine the basic premium amount), but also about the whole distribution of the aggregate claim for the year.

We model the aggregate claim in two independent parts – the number of claims that occur per year and the size of each claim. The sum of these individual claims is the aggregate claim. We call this the aggregate risk model.

In this dissertation, we examine Bayesian methods as they can be applied to forecasting future aggregate claims using the aggregate risk model. Bayesian methods are particularly useful in the field of actuarial science, as they enable us to learn about the whole distribution of a chosen quantity and also to include the many levels of variation involved in the models we use. Under the Bayesian approach, we assume that all parameters in the model are random and have suitable prior distributions, which we update with the data. This approach reflects the uncertainty inherent in estimating these unknown values and the posterior distributions derived give us an indication of how reliable the inferences are. With Bayesian methodology we are able to incorporate additional data into the analysis, as it becomes available, to refresh the inference made and increase the reliability of the estimates. Once we have found a suitable model and learnt about the posterior distributions of the parameters, we may use the information gathered to give the distribution of aggregate claims.

In the first chapter, we examine the theoretical background of the main tools of our analysis: Bayesian analysis and Monte Carlo simulation techniques for inference. We take a brief look at the history of Bayesian results and see how this approach has

combined with computing methods into techniques often used nowadays in actuarial science. In Chapter 2, we deal with a specific example of aggregate claims and see how we can apply the techniques in this instance to learn about the parameters of a variety of Compound Poisson models. We compare the results to those derived from classical methods. Chapter 3 uses the inferences made in the second chapter to make predictions about future aggregate claims and to examine the predictive distribution of the aggregate claim, which could be used by insurers to set reserves for policies. The final chapter highlights the conclusions of this analysis.

1 THEORETICAL BACKGROUND

1.1 Bayesian Inference

We discuss the problem of estimating an unknown parameter θ (where θ may be multivariate), given data $\underline{y}$ from an experiment, which we consider as a random sample from a probability model depending on θ.

Using the classical or frequentist approach, the form of the experiment is to take a sample of size n (where n is large) from the distribution, and we think of θ as a fixed, but unknown parameter. However, in many circumstances it is not possible to sample a large number of data points, yet we may still want to make inference about θ. An example would be to estimate the proportion of people who vote for one party in an election by using the results of polls. We have only a small number of data points in this case so the classical model may give us a misleading answer, particularly if the poll is not a representative sample. We are also unable to use any previous knowledge about likely values of θ within the classical framework.

An alternative approach is to use Bayesian inference. We now regard θ as a random variable. The information and beliefs of the investigator about the possible values of θ prior to observing the data are summarised in the prior distribution $\pi(\theta)$. Our aim is to combine this with the data $\underline{y}$ to obtain the posterior distribution $\pi(\theta \mid \underline{y})$. Bayes' theorem (a simple result from conditional probability – see [1] for details) tells us that:

$$\pi\left(\theta \mid \underline{y}\right) = \frac{\pi(\theta) f\left(\underline{y} \mid \theta\right)}{\sum_i \pi(\theta_i) f\left(\underline{y} \mid \theta_i\right)}, \tag{1}$$

where $f(\underline{y} \mid \theta)$ is the likelihood function. We can generalise this to the continuous case by replacing the sum by an integral. The denominator is a constant with respect to θ, so we may simplify the expression to:

$$\pi\left(\theta \mid \underline{y}\right) \propto \pi(\theta) f\left(\underline{y} \mid \theta\right). \tag{2}$$

By appealing to the likelihood principle (see [2] for a detailed discussion), we see that this simplification loses none of the information provided by the data.

1.2 A Brief History of Bayesian Methods and Choice of Priors

The term Bayesian analysis comes from the Rev. Thomas Bayes, whose paper "An essay towards solving a problem in the doctrine of chances" [3] was published posthumously in 1763. This paper had little influence at the time, and Laplace in fact developed independently the roots of what is now termed Bayesian statistics, in his book [4]. He attempted to justify the use of uniform prior densities in cases where we have no prior information about the distribution of the parameter in question. Bayes himself chose the prior density by using physical reasoning, which works well for some cases, but not for a general problem.

Jeffreys wrote a "Theory of Probability" [5], and in it expounded his theory on Bayesian methods, including a general formula for the Jeffreys prior:

$$\pi(\theta) \propto \sqrt{I(\theta)}, \qquad (3)$$

where $I(\theta) = -E_{\underline{Y}|\theta}\left[\partial^2 \log L(\theta \mid \underline{Y}) / \partial\theta^2\right]$ is the Fisher Information.

This prior is designed to be as uninformative as possible, to minimise the possibility that it influences the data. It is very difficult to find a truly uninformative prior, due to the requirement that any transformation of the parameter needs to have an equally uninformative prior. For example, if we take the Uniform [0,1] distribution as an uninformative prior for a proportion p (which appears to be a natural choice, as we attribute an equal probability to each possible value of p), then we would like $q = p^2$ to have the same prior density, which clearly does not hold. One problem that may occur with Jeffreys priors, and indeed also with uniform priors over an infinite range, is that they cannot be normalised, and are hence improper priors. Provided the posterior density is a proper distribution however, most Bayesians accept their use. The theoretical justification for uniform priors is still contentious, but they remain a widely used form of prior in practice.

Some theoreticians, including B. de Finetti, L.J. Savage and D.V. Lindley, rejected frequentist methods, as well as the use of the uninformative prior, in favour of subjective priors. In some situations, subjective priors are appropriate, but they claimed that they should always be used. The consequence of this approach is that different statisticians may obtain different results from the same data, which is undesirable.

Until recently, most statisticians used classical methods, developed by R.A. Fisher and Neyman and Pearson (see [6]). However, with the advent of computing, Bayesian methods are becoming more prevalent due to the use of increasing computational power to evaluate complicated posterior distributions by numerical integration or Monte Carlo methods. Often we use conjugate priors (priors for which the associated posterior is a member of the same family of distributions) in these cases, to simplify the calculations. This also makes it easier to obtain a closed form expression for the estimator (see [1] for a more detailed discussion of conjugate priors). Before the arrival of modern computational power, use of conjugate priors was the only technique available to apply Bayesian methods in a way that was mathematically tractable. With the aid of technology, we are able to implement Bayesian methods far more widely than before.

1.3 Markov Chain Monte Carlo methods

To adapt Bayesian methods for computing, we need to find a technique to generate samples from our posterior distribution $\pi(\underline{\theta} \mid \underline{y})$, or more generally from some target distribution π, since we cannot sample directly. One possible way to solve this problem is to use Markov Chain Monte Carlo (MCMC) methods as follows. We aim to construct a Markov chain with state space $\Theta = \{\theta_i\}$ and equilibrium distribution equal to our target distribution, from which we can simulate. We then run the chain and use the simulated values to draw inferences and conclusions.

The general procedure we use is the Metropolis-Hastings algorithm (for more details of which refer to [7]). We describe the process for discrete univariate θ. It can be generalised easily for any θ. Let q_{ik} be an arbitrary transition probability function, so if the chain is currently at state θ_i, we choose a candidate state $\theta^* = \theta_k$ with probability q_{ik}. The next step introduces another level of randomisation. We decide whether to accept or reject this candidate state, so we set the next state to be θ^* with probability p_{acc} or θ_i with probability $1 - p_{acc}$, where we define p_{acc} by:

$$p_{acc} = \min\left\{1, \frac{\pi(\theta_k)q_{ki}}{\pi(\theta_i)q_{ik}}\right\}.$$

It can be proved that this yields a Markov chain with equilibrium distribution π, provided we choose q to be irreducible and aperiodic. The choice of q, within these restrictions, does not affect the value to which the chain converges but can reduce the efficiency of

the method if it forces p_{acc} to be low in value. Therefore, we may need to take care in choosing q. Since the target distribution only enters the calculations in the form of ratios, it makes the Metropolis-Hastings algorithm particularly useful for Bayesian applications, where we often only know the values of π up to a constant of proportionality.

For multivariate $\underline{\theta}$, it is often more efficient to update each of the components θ_i of $\underline{\theta}$ individually. This is referred to as single-component Metropolis-Hastings. A special case of this is Gibbs sampling (developed by Geman and Geman in [8]). For this method, we take q to be the conditional posterior distribution $\pi\left(\theta_i \mid \underline{y}, \theta_j : j \neq i\right)$ to update the i^{th} component. This is easily identified from $\pi(\underline{\theta} \mid \underline{y})$. We assume that we are able to sample from q. This method ensures that the value of p_{acc} is always 1. To implement the routine, we follow the following algorithm:

- Choose arbitrary starting values for all the components $\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_n^{(0)}$.
- For $1 \leq k \leq n$ and $t = 1$, iterate by drawing $\theta_k^{(t)}$ from

  $$\pi\left(\theta_k \mid \underline{x}, \theta_1^{(t)}, ..., \theta_{k-1}^{(t)}, \theta_{k+1}^{(t-1)}, ..., \theta_n^{(t-1)}\right).$$
- Repeat step 2 for $t = 2, 3, \ldots$.

This results in a sequence $\left(\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_n^{(t)}\right)$. By repeating many times we obtain random samples from a joint distribution that tends to the target distribution as t increases. Usually several thousand iterations are required. To increase the efficiency of this method, we often use conjugate priors to generate the θ_i.

We often run multiple copies of the Markov chain to obtain a larger random sample, but we may instead run one longer chain. Doing this means that successive samples are likely to be correlated, so in order to obtain a random sample we must thin the chain. We therefore record only every m^{th} state output by the chain. If we choose a sufficiently large value of m then consecutive samples will be approximately independent. If the value of m required for this is small, then we say that the mixing qualities of the chain are good.

When using MCMC methods, we must verify that the chain has indeed reached the equilibrium distribution before we take samples to use for inference. This is because our choice of starting values is arbitrary, and particularly for slow-mixing chains there may be a long period required for "burn-in". The burn-in depends on the speed of convergence of the chain and the starting position, as the initial distribution may heavily

influence the simulation for a long period. There are several methods to determine whether a chain is converging adequately. For a discussion, refer to [9]. We discuss this problem in more detail in Section 2.4.

1.4 Comparison of Classical and Bayesian Methods

The two theories of Statistics often complement each other, as they have different strengths and weaknesses. Some theoreticians do have a strong preference for one of the approaches. A point of criticism of the Bayesian method is the subjectivity inherent in the choice of prior. As mentioned in Section 1.2, there are several different ways in which one may choose a prior, some of which can be justified better than others. In practice, the distinction between the two methods may be less important. Asymptotic posterior normality tells us that, under appropriate conditions, our beliefs about $\underline{\theta}$ coincide with the "true" parameter value as the amount of data observed increases. For a proof of asymptotic posterior normality, see [1]. This is the Bayesian analogue of the Strong Law of Large Numbers. When the samples are large, the data overwhelms the information given by the prior for the Bayesian method and leads to answers very similar to those resulting from the classical method. This also reduces the importance of selecting the "correct" prior.

The classical method may lead to certain inconsistencies. The following example highlights one such paradox:
Suppose we observe the following data Y, and we wish to test the null hypothesis that $Y \sim Poisson(1)$.

Table 1.4.1 Observed values of data Y and expected frequencies under the null hypothesis $Y \sim Poisson(1)$. Source: own calculations.

y	0	1	2	3 or more
Observed Frequency f_y	31	42	23	4
Expected Frequency e_y	36.79	36.79	18.39	8.03

The value of the Pearson χ^2 statistic for these data is 4.9315, with 3 degrees of freedom, giving a p-value of 0.177, which would lead us to accept the null hypothesis at the 10% level. However, if we hypothesise instead that $Y \sim Poisson(\lambda)$ with λ unknown, we need to estimate λ to proceed. Using maximum likelihood estimation gives $\hat{\lambda} = 1$, therefore giving the same expected frequencies and χ^2 value as before. We now have only 2

degrees of freedom, as we were required to estimate a parameter, so the p-value is thus 0.084, leading to rejection of the null hypothesis at the 10% level. To summarise, the second version of this test has a smaller p-value than the first, despite having a weaker null hypothesis. This indicates that there is more evidence against the weaker null hypothesis than the stronger in this example, which is inconsistent.

The following outlines the main arguments in favour of and against each method.

Classical approach:

i) Good for standard examples using common distributions and usually quick and easy to compute the solution.
ii) Asymptotic Maximum Likelihood theory is useful for large samples.
iii) Can give rise to paradoxes and subjectivity, for example with p-values and confidence intervals, as shown in the example above.
iv) Predictions are based on long-term repeated events.

Bayesian approach:

i) Intuitive method to include prior knowledge (if it is available and can be expressed probabilistically).
ii) Logical way to collate data from several experiments.
iii) Intellectually coherent, as inconsistencies arise only from different choices of prior distribution.
iv) Easy to judge accuracy of any particular inference and easy to interpret results.
v) Choice of prior is subjective, and difficult to find a truly uninformative prior. Rarely have sufficiently comprehensive prior information.
vi) Can be computationally heavy.

To summarise, the Bayesian model requires more input (in the form of additional assumptions about the prior) than the classical model, but it is often easier to interpret the results, as the inference about θ is given in the form of a distribution.

1.5 Bayesian Methods in Actuarial Science

Actuarial science uses Probability Theory and Statistics to analyse data about past events in order to make predictions regarding what may happen in the future, particularly in relation to financial matters, such as insurance, pensions or investments.

We restrict our consideration to events that we can measure in monetary terms, such as the destruction or loss of some item or property, which by assumption occur independently and at random.

In risk theory, we aim to learn about the loss process underlying this given series of events, for example claims on a motor insurance portfolio. We use the information gathered on the number and severity of claims in previous years to provide inference about the distribution of the loss process, and thereby charge a suitable premium for the insurance policies. Insurance claims, as we expect, are uncertain in both frequency and amount. The purpose of insurance is to make the costs to the insured parties more stable over time. To this end, insurers sell large numbers of policies for a set premium, with the promise that they will pay for losses sustained. By pooling all the risks, the insurer reduces the risk of loss to itself, and to reduce the risk still further, may additionally take out reinsurance for individual or aggregate claims over a certain amount.

The insurer is interested in particular in the aggregate claim S_t – the sum of all the claims it receives in year t – so it may ensure it has sufficient funds to pay for the claims it receives. We assume that S_t has the same distribution for all values of t. We call the funds that the insurer sets aside for this purpose reserves. We can find the density function of S_t using classical approaches, such as convolution methods, but this can be a very complicated way to proceed and may require lengthy numerical methods for integration. We could instead find point estimates of S_t, but this can be restrictive. An alternative is to use Bayesian methods, combined with MCMC simulation to perform the calculations required for inference. The form of MCMC that is most used for actuarial analyses is the Gibbs sampler [10]. MCMC provides an estimate of the entire posterior density of S_t, from which we may calculate whatever statistics may be of interest to us. We may also predict the future aggregate claim S through use of predictive distributions. For more information about predictive distributions, see Section 3.

Bayes methods impact on a number of other areas of actuarial investigations, including generalised linear models and credibility theory in particular, but in this project we focus solely on aggregate claims.

2 AGGREGATE CLAIMS MODEL

2.1 Assumptions and Design of Model

Throughout this project, we aim to analyse data taken from [11]. The data are all single loss amounts greater than 1.5 million for the last 5 years (indexed for inflation) from a motor portfolio, and are presented in Table 2.1.1.

Table 2.1.1 Claims (in millions) exceeding 1.5 million over five years from a motor portfolio. Source: Rytgaard (1990).

Year number	1	2	3	4	5
Claims	2.495	1.985	3.215	-	19.180
	2.120	1.810	2.105	-	1.915
	2.095	1.625	1.765	-	1.790
	1.700	-	1.715	-	1.755
	1.650	-	-	-	-

The lower bound of 1.5 million refers to the retention of this excess-of-loss policy. Insurers often choose to limit the amount of claims they are required to pay themselves by taking out reinsurance. This ensures that the variation experienced by the insurer is not too large. Reinsurance means that if the claim y is above a certain amount d, called the retention or retained part, then the insurer pays only d to the insured party and the excess $y - d$ is paid by the reinsurer. Insurers use a similar process, with the retention known instead as a deductible, to prevent small claims being made. They do this to reduce their costs, as processing small claims can be time-consuming and costly relative to the amount insured. Sometimes, they use a franchise deductible, which is very similar to a straight deductible, but the insurer pays the full amount y provided y exceeds d.

To model this data, we shall use a Compound Poisson Distribution, as described in [12]. Let N_t be the number of claims in year t, and let $Y_{i,t}$ be the amount of the i^{th} claim in year t, with $i = 1, 2, \ldots, N_t$. Let $S_t = Y_{1,t} + Y_{2,t} + \ldots + Y_{N_t,t}$ be the aggregate claim amount in year t. We assume that:

- $N_t \sim Poisson(\theta)$, for some $0 < \theta < \infty$, are iid for all t.
- $Y_{i,t}$ are iid for all i, t.
- N_t and $Y_{i,t}$ are independent for all i, t.

Then S_t has the general form of the Compound Poisson Distribution. The Poisson distribution is a natural one to use to model the frequency of claims, as we assume that claims occur at a constant rate in time at random. The distribution of $Y_{i,t}$ for this situation must have the form of a loss distribution. A loss distribution is one that only takes positive values and has long tails in order to allow large claims (see [13] for more information about loss distributions). We shall initially use a Poisson-Pareto model; that is we assume that:

$$Y_{i,t} \sim Pareto(\alpha, \beta), \qquad \alpha > 0, \quad 0 < \beta < y. \tag{4}$$

The Pareto distribution is often used to model insurance claims, as it has a fat tail, with the density tending to zero like $y^{-\alpha}$ as y tends to infinity. The form of the Pareto distribution that we are using in this analysis is given in Appendix 1, and is particularly suitable to model claims with a deductible, as we are able to choose the lower limit of observations. We will consider several other distributions as alternative models in Section 2.3.

2.2 Running the MCMC in WinBUGS

We will use Bayesian MCMC methods to analyse these data and investigate the distribution of S_t, treating α, β and θ as random parameters in this Poisson-Pareto model. Since we know very little about the true values of these parameters, we will use vague Gamma priors, as follows:

$\alpha \sim Gamma(1, 0.0001)$,

$\beta \sim Gamma(1, 0.0001)$, with the restriction that $0 < \beta < \min\{y_{i,t}\}=1.625$,

$\theta \sim Gamma(1, 0.0001)$.

The restriction on the distribution of β is due to the requirement that $\beta < y$ in the specification of the Pareto model (in equation (4)). These priors each have a variance of 10^8, so should not influence the posterior distribution much, and the Gamma distribution is a conjugate prior for the Poisson and Pareto distributions, as we will verify later. A large prior variance is indicative of a vague distribution and therefore reflects our relative ignorance about the true parameters. We will consider alternative priors in Section 2.5. We derive the posterior distributions as follows (using equation 2 and the same notation as in Chapter 1).

$$\pi(\alpha,\beta \mid \underline{y}) \propto \pi(\alpha)\pi(\beta) f(\underline{y} \mid \alpha,\beta)$$
$$\propto e^{-0.0001\alpha} e^{-0.0001\beta} \alpha^{16} \beta^{16\alpha} \left(\prod_{i=1}^{16} y_i \right)^{-(\alpha+1)}$$
$$\propto \alpha^{16} \exp\left\{ -\left(0.0001 + \sum_{i=1}^{16} \ln y_i \right) \alpha \right\} . \beta^{16\alpha} \exp\{-0.0001\beta\}.$$

This gives the following full conditional posterior distributions for α and β:

$$\pi(\alpha \mid \beta, \underline{y}) \propto \alpha^{16} \exp\left\{ -\left(0.0001 - 16\ln\beta + \sum_{i=1}^{16} \ln y_i \right) \alpha \right\},$$
$$\pi(\beta \mid \alpha, y) \propto \beta^{16\alpha} \exp\{-0.0001\beta\},$$

which implies that:

$$\alpha \mid \beta, \underline{y} \sim Gamma\left(17, \sum_{i=1}^{16} \ln y_i - 16\ln\beta + 0.0001 \right), \quad (5)$$

$$\beta \mid \alpha, \underline{y} \sim Gamma(16\alpha + 1,\ 0.0001). \quad (6)$$

Similarly, for θ we obtain the following posterior distribution:

$$\pi(\theta \mid \underline{n}) \propto \pi(\theta) f(\underline{n} \mid \theta)$$
$$\propto e^{-0.0001\theta} e^{-5\theta} \theta^{\sum n_i}$$
$$\propto e^{-5.0001\theta} \theta^{\sum n_i}$$

which gives us that:

$$\theta \mid \underline{n} \sim Gamma\left(\sum_{i=1}^{5} n_i + 1, 5.0001 \right). \quad (7)$$

We see that (5), (6) and (7) are all Gamma distributions, as Gamma priors are conjugate for the Poisson-Pareto model. We will be using a program called WinBUGS to analyse the distributions underlying the data (for more information on WinBUGS, see [14]). This implements the Gibbs sampling technique in the case where we use conjugate priors (described in Section 1.3) to obtain Bayesian inference. When the priors are not conjugate, WinBUGS uses other sampling techniques. To use WinBUGS, we do not need to specify these posterior distributions, but we must input the model and the data and we need to give some initial values for α, β and θ. We use well-dispersed values to initialise the model, running three chains in parallel to check for convergence. If the three chains start from differing places yet converge to the same parameter values, then that is a good indication that the chains have reached the equilibrium distribution. We list the initial values that we have used in this model in Table 2.2.1.

Table 2.2.1 Initial values used for the parameters for the three Markov Chains in the Poisson-Pareto Model.

Chain Number	1	2	3
α	0.00001	100000	3.076
β	0.00001	1	1.625
θ	0.00001	100000	3.2

The values for the third chain are the Maximum Likelihood Estimates (MLEs) of the parameters, as we would expect the posterior mean values to be close to the MLEs. The computer code to be used in WinBUGS for this model is in Appendix 2.

From examining the various convergence diagnostics, such as Figure 2.2.1 (a segment from the trace history of the values of α, which shows the three chains have mixed well), we can see that the chain has converged by the 20000th iteration. More details about the convergence diagnostics are contained in Section 2.4.

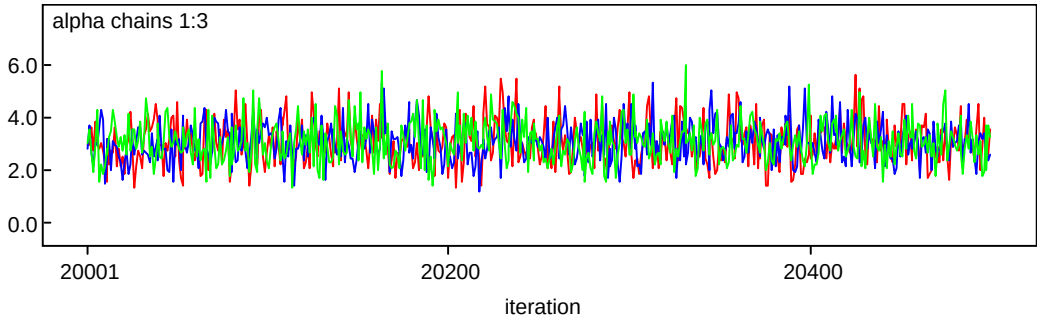

Figure 2.2.1 Trace history of the value of α, from the 20000th iteration, in the Poisson-Pareto model. Source: own calculations.

Combining the results from the next 30000 iterations yields the results in Table 2.2.2.

Table 2.2.2 Posterior estimates of the Poisson-Pareto model parameters from 30000 iterations of the three Markov Chains combined, after a burn-in of 20000. Source: own calculations.

	Posterior Mean	Posterior Standard Deviation	95% Bayesian credible interval
α	3.083	0.7686	(1.759, 4.755)
β	1.591	0.03489	(1.497, 1.624)
θ	3.400	0.8217	(1.983, 5.202)
Average Claim	2.506	1.333	(2.022, 3.648)

We can see that these values are quite close to the MLEs. The sample mean for Y is 3.058, which lies comfortably within the 95% Bayesian credible interval for the average claim amount (which is the mean single claim loss calculated at each iteration from the observed values of α and β). The sample mean is larger than the posterior mean, because of the large value of 19.180 in the data set, which is an outlier and skews the results. By comparison, the mean of Y calculated using the posterior estimates of α and β is 2.355, which is close to the median value of the posterior mean, 2.381. The 95% Bayesian credible intervals for the parameters are quite wide. This reflects the relatively small amount of data we have available. The credible interval for β contains the whole range we would expect, since we know that β is less than 1.625 (as that is the minimum value of y observed) and we know the data points have 1.5 as the lower limit, so we anticipate that $\beta \geq 1.5$. Figure 2.2.2 illustrates the posterior distributions of these variables.

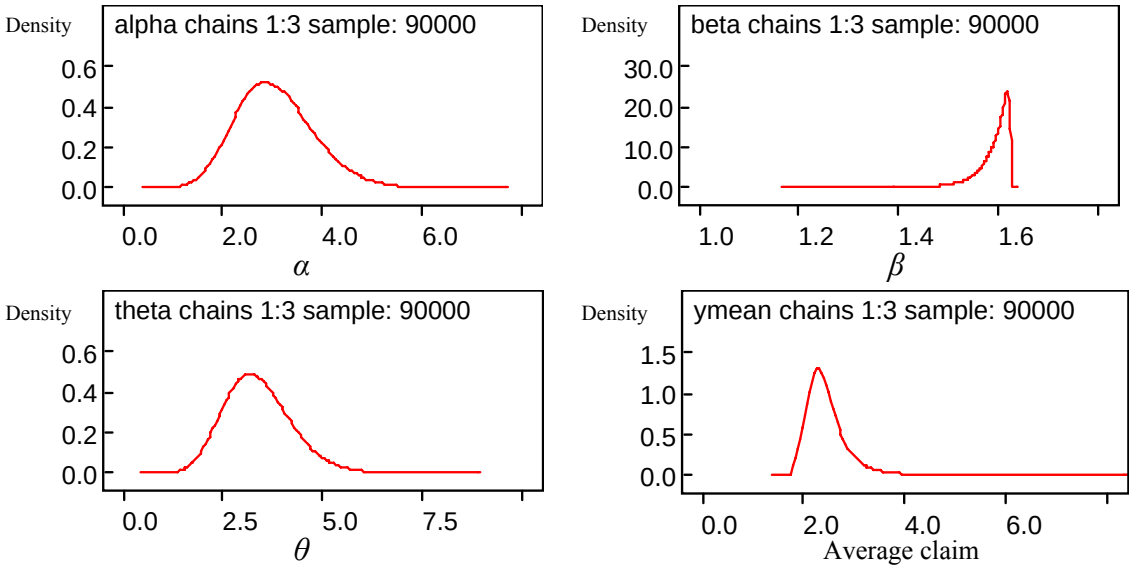

Figure 2.2.2 Posterior densities for α, β, θ and the average claim amount.
Source: own calculations.

We can see in Figure 2.2.2 that the posterior densities of the parameters do look like Gamma distributions (with the density for β truncated at 1.625). This is due to the conjugacy of the priors. The posterior mean density can have some very large values (albeit with very small probability), corresponding to values of α near to 1 in value.

We can adjust the Pareto distribution slightly in the model, for example by using only one parameter (and hence setting $\beta = 1.5$, as we know that the lower bound for the data is in fact 1.5). This model gives the Bayesian estimate of α as being equal to 2.623, and

that of the average claim amount as being equal to 2.668, which is similar to before. Figure 2.2.3 compares the density of Y using the two different Pareto distributions with the data.

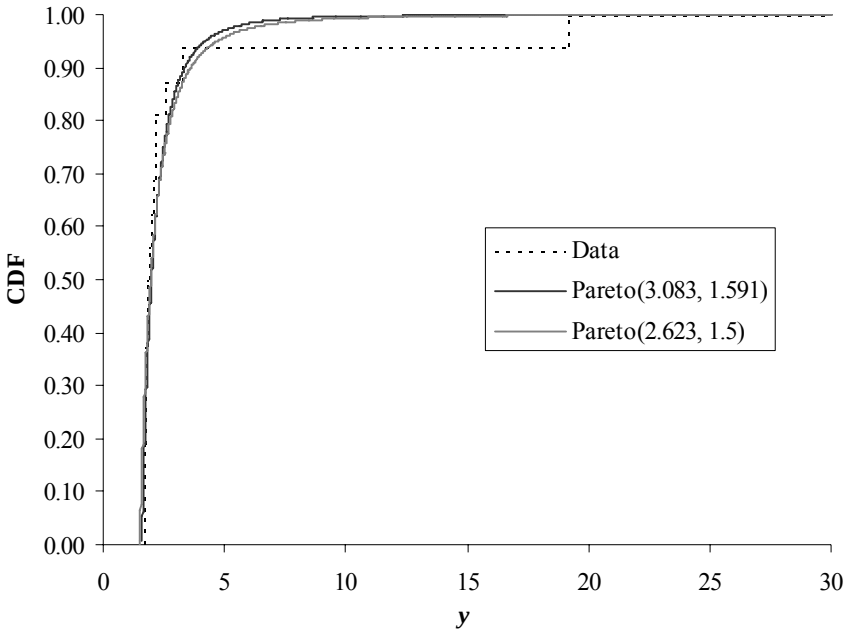

Figure 2.2.3 Graph of cumulated data y_i with cumulative density functions of *Pareto*(3.083, 1.591) and *Pareto*(2.623, 1.500) superimposed for comparison. Source: own calculations.

We can see that there is little difference between the two models, although the one with $\beta = 1.5$ fixed has a slightly fatter tail and consequently would have more chance of large values. It is difficult to tell which model fits the data best from this graph, however in this project we prefer to estimate the β parameter, as this provides a more flexible model. Both Pareto models have a very close fit to the data.

Similarly, we can compare the *Poisson*(3.4) distribution to the data n_t. The comparison is in Figure 2.2.4. This seems to fit fairly well. The data set is obviously small, but it follows the general shape of this Poisson distribution. Despite the fact that the data is effectively censored, with points less than 1.5 million in value not being recorded, the Poisson distribution still fits the data well. This is due to the splitting property of Poisson processes, which means that, for example, if the occurrence of claims in any year is *Poisson*(λ), and if the probability of a claim being greater that 1.5 million is p, then the occurrence of claims greater than 1.5 million in any year is *Poisson*(λp). This is one benefit of using the Poisson distribution to model these data.

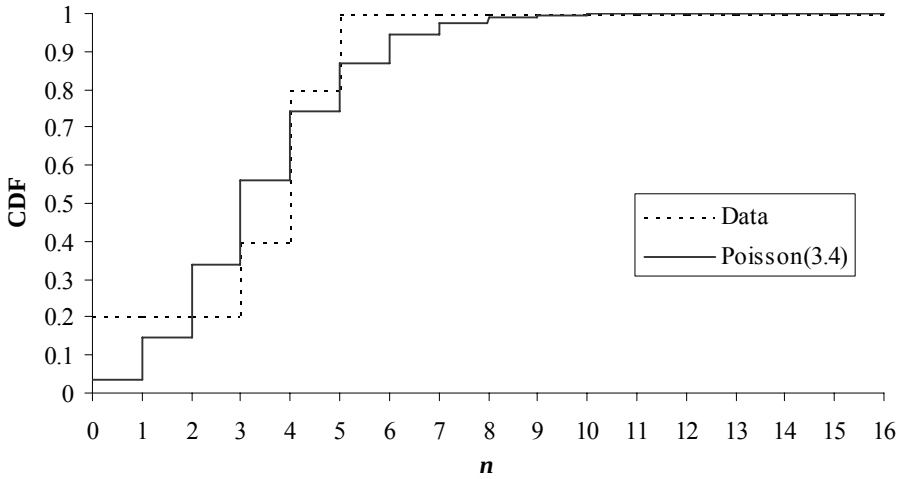

Figure 2.2.4 Graph of cumulated data n_t with density *Poisson*(3.4) superimposed for comparison. Source: own calculations.

2.3 Alternative Models for the Loss Distribution

There are many other loss distributions that we may consider to model the individual claim amount Y for these data. Some of the simplest to use (since they contain at most two parameters to be estimated) are the Gamma, Exponential (a one parameter version of the Gamma), Lognormal and Weibull. There are of course many other variations we could also try, such as the Burr (which is closely related to the Pareto distribution) or Loggamma distributions, but we shall only examine these four in this project. The method is the same as for the Pareto model and we will use the same vague Gamma priors for these distributions too (except for the parameter α of the Lognormal distribution, which can take any real value, and for which we shall use a vague Normal prior). A slight modification to these distributions is to use truncated forms of the densities. Since all the observations must be greater than 1.5 due to the form of the data, we may be able to ensure a better fit if we truncate the distributions before fitting them to the model, rather than fitting a distribution and then truncating it. We truncate the distribution by restricting the support to $y > 1.5$, rather than y strictly positive, and then normalising the density function by dividing it by the probability that $Y > 1.5$. Therefore, the general form of a density truncated to 1.5 (where the cumulative distribution function is denoted $F(y)$ and the probability density function is $f(y)$) is:

$$h(y) = \frac{f(y)}{1 - F(1.5)} \tag{8}$$

The forms of the densities used for both varieties of these distributions are displayed in Appendix 1. As the densities of the truncated distributions are not supported in

WinBUGS, we used a technique that specifies a likelihood proportional to that of the truncated model. An example of the code used is in Appendix 3. For the Truncated Gamma distribution, we had to use the Incomplete Gamma function. To do this, we calculated an approximation based on the infinite series expression in equation 9.

$$\Gamma(\alpha;x) = e^{-x}x^{\alpha}\sum_{n=0}^{\infty}\frac{\Gamma(\alpha)}{\Gamma(\alpha+1+n)}x^{n} \qquad (9)$$

We do not attempt to truncate the Lognormal distribution, because the equation for its cumulative density is very complicated to use when programming.

The posterior mean (μ), standard deviation (σ) and 95% equal-tailed Bayesian credible intervals of the parameters are shown in Table 2.3.1.

Table 2.3.1 Posterior parameter values for the Exponential, Gamma, Weibull and Lognormal distributions. Source: own calculations.

Claim size model	α			β		
	μ	σ	95% Interval	μ	σ	95% Interval
Exponential(α)	0.348	0.084	(0.203, 0.532)	N/A	N/A	N/A
Trunc. Exponential(α)	0.682	0.165	(0.398, 1.043)	N/A	N/A	N/A
Gamma(α, β)	2.074	0.625	(1.047, 3.492)	0.699	0.236	(0.314, 1.226)
Trunc. Gamma(α, β)	0.328	0.318	(0.009, 1.178)	0.514	0.173	(0.231, 0.904)
Weibull(α, β)	0.277	0.108	(0.113, 0.531)	0.920	0.683	(0.899, 1.465)
Trunc. Weibull(α, β)	0.251	0.242	(0.024, 0.908)	0.320	0.179	(0.054, 0.726)
Lognormal(α, β)	0.810	0.15	(0.511, 1.107)	0.357	0.140	(0.177, 0.710)

By comparing the parameter values for the Truncated Gamma, Truncated Weibull and also the original form of the Pareto (from Table 2.2.2) to the values in [15], we can see that these estimates are close to the posterior values found by Pai. Figure 2.3.1 displays the cumulative densities of the distributions detailed in Table 2.3.1, compared to the cumulated data. From examining the graphs, we can see that in general the truncated distributions do have a better fit than the standard distributions, as we would expect. The Gamma distribution fits better than the Exponential, as it is a more general function. We could probably find distributions that have a superior fit than these, by investigating those with three parameters, but we restrict ourselves to two for ease of calculation. From this set of distributions, it appears that any of the truncated distributions may be considered optimal, but the fit of all three is inferior to that of the Pareto model. The

standard Weibull distribution has a very different fit to the others, since it appears to model the smaller claims better, but has too thin a tail to model this loss distribution. If we wished, we could test the fit to the data more formally by using Bayes factors, comparing deviances or performing χ^2 tests after discretising the distributions.

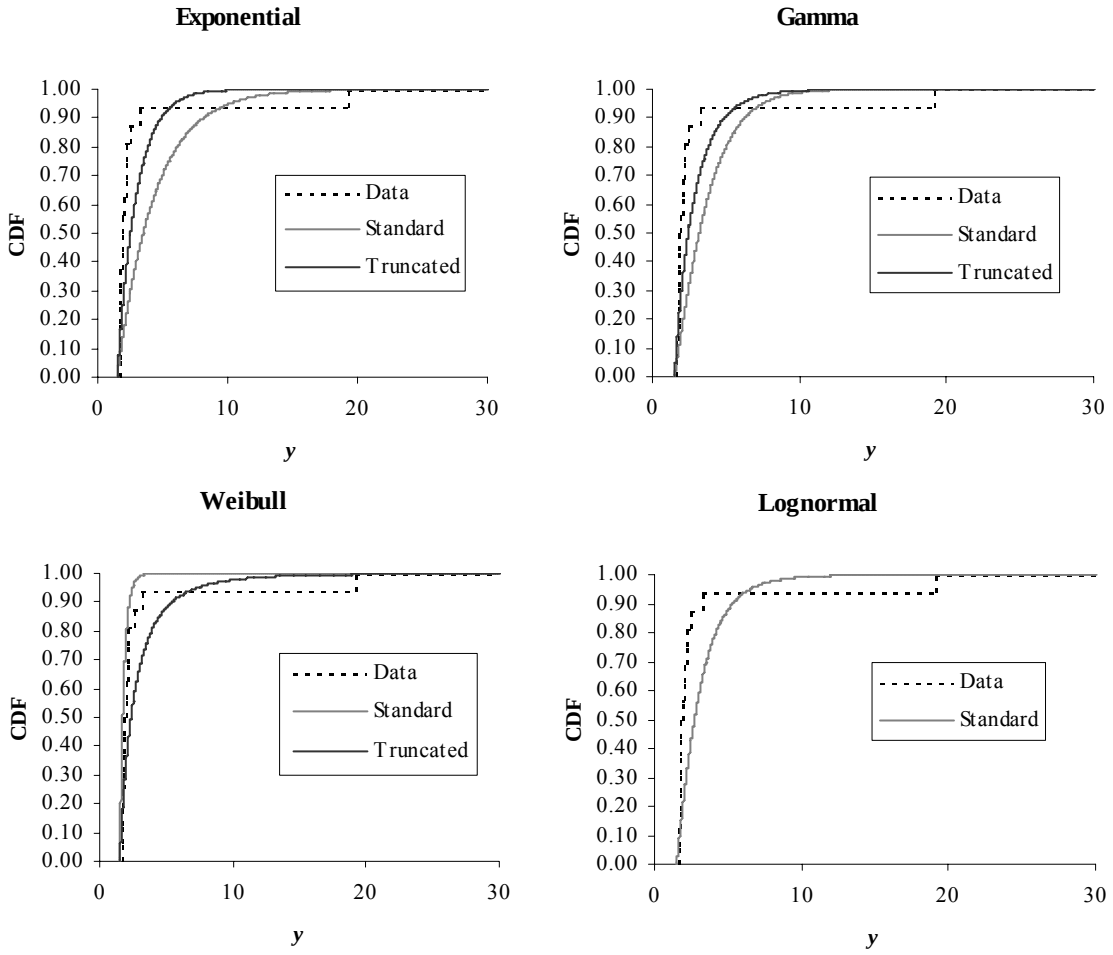

Figure 2.3.1 Cumulative density functions of fitted loss distributions compared to data. Source: own calculations.

To summarise, the Pareto appears to model the claim size substantially better than the other distributions considered. This supports the fact that this distribution is often used as a loss distribution. The other distributions examined here do fit the data reasonably well, with the two-parameter distributions generally fitting better than the single parameter Exponential distribution, although the fit is improved in all cases if we aim to fit a truncated distribution to the data. This is because we know that all the data points listed are greater than 1.5 million, as they are claims from a policy with retention or deductible of 1.5 million. If we had the full set of data points, we would be able to provide a superior model, and then truncate that to provide inference about the aggregate claims. Unfortunately, it is often the case that data are censored in some way.

In addition to comparing the cumulative distribution functions, we can also compare the distribution of the average claim using different models. In Figure 2.3.2, we show the posterior densities of the average claim for three different models: the two-parameter Pareto, the Truncated Exponential and the Lognormal.

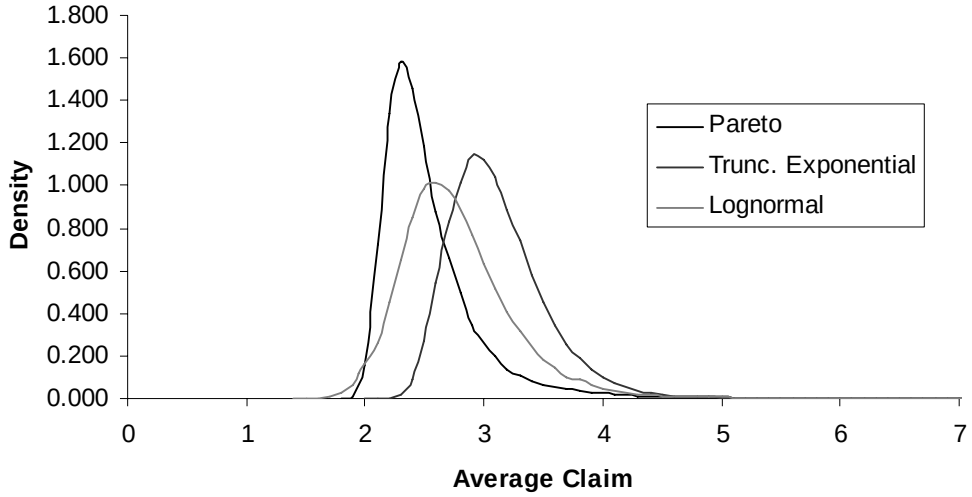

Figure 2.3.2 Posterior densities of the average claim amount for three different models. Source: own calculations.

We chose these models as they have a reasonable fit to the data, and the calculation of the average claim amount is straightforward. We see that the distribution of the posterior mean does vary across the distributions, although the values are all quite similar. The average claim amount using the Pareto model has a smaller posterior mean, as the claim size distribution with the Pareto has a thinner tail than the others do. The models will thus give slightly different results.

2.4 Convergence Issues

As mentioned before, it is imperative that we ensure convergence of the Markov Chains used before we take results from them. In WinBUGS, there are several different methods to check for convergence. Here, we describe some of those methods.

We first need to estimate the length of the burn-in period, before we can take a sample from the converged chain. To estimate the burn-in, we may look at the trace of the chains to see whether they have mixed well. As an example for this section, we shall use the Gamma model from Section 2.3. All graphs come from the output of WinBUGS. Figure 2.4.1 shows a segment of the trace history of parameter α for iterations from 10000. As we can see, the chains appear to have mixed well by this stage.

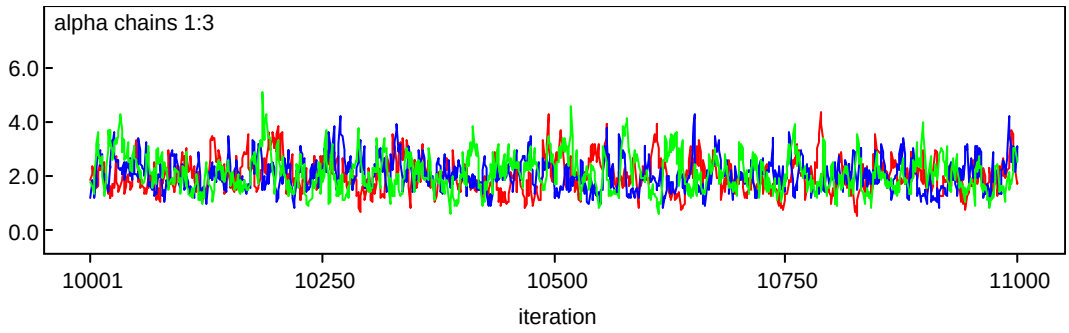

Figure 2.4.1 Trace history for Gamma model, after burn-in of 10000.

A more precise method to verify convergence is to look at the Gelman-Rubin statistic (described in [16]). WinBUGS uses a modified form of the original Gelman-Rubin statistic, called the Brooks-Gelman-Rubin (or BGR) test. To perform this test, we need to run two or more chains in parallel, with initial values over-dispersed relative to the true posterior. The BGR test compares the variances within and between the chains. In the plots of the BGR statistic in Figure 2.4.2, the lower two lines represent the within and between chain variations respectively and the upper line is the ratio of the between and within chain variations. When the lower two lines stabilise (in these diagrams at approximately 10000 iterations) and the upper line converges to 1, then we accept that the chain has converged.

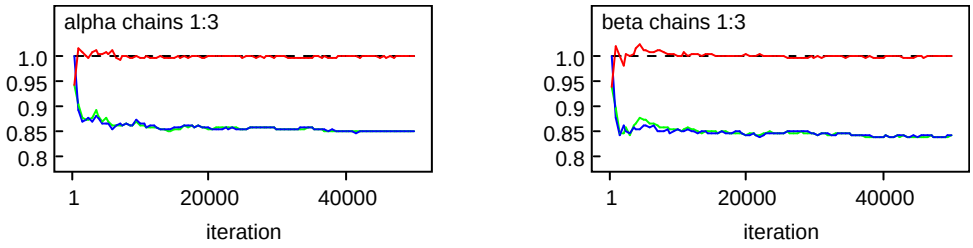

Figure 2.4.2 BGR diagrams for α and β.

There are several other convergence diagnostic tests that we may use, including Geweke (1992), Raftery & Lewis (1992) and Heidelberger & Welch (1983). For more details on these and on the BGR test, see [17].

To obtain an independent sample, we may wish to thin the converged chain, so that the correlation between successive states is small. The auto-correlation functions for parameters α and β for the Gamma model are shown in Figure 2.4.3.

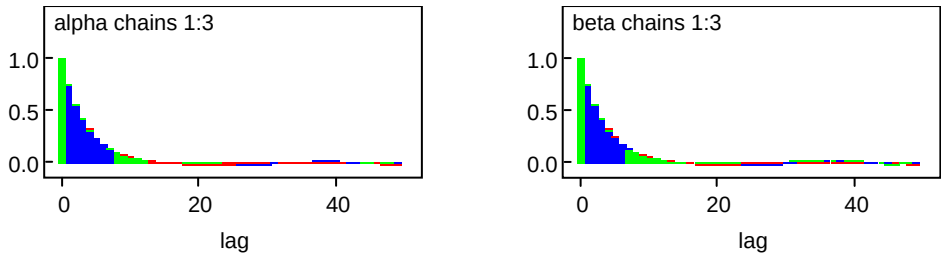

Figure 2.4.3 Auto-correlation functions for α and β, before thinning.

Figure 2.4.3 shows us that α and β both have some auto-correlation. This is undesirable, as we would like the samples to be approximately independent, so we thin the chain by using only every m^{th} value from the output as described in Section 1.3. Thinning the output by 10 reduces the auto-correlation to less than 0.07 for all lags greater than or equal to one. This is an acceptably low level of auto-correlation and thus we accept this thinned chain as representing an independent random sample from the required distribution. Thinning the sample does have potential problems. It does create an approximately independent sample, although it may also disguise flaws inherent in the model and does not in itself improve the rate of convergence. It also means that in order to obtain a sample of a given size, we need to perform a considerably larger number of iterations, which may be costly in terms of computing time. The models we have been examining in this project do not take long to run, and therefore this is not a problem here.

One final, and straightforward, way to check for convergence is to look at the Monte Carlo standard error of the posterior mean, which can be obtained from WinBUGS. The method used to calculate it is the batch means method, described in [9]. We want these values to be less than 5% of the sample standard deviation. In all our examples, it is in fact less than 1%.

2.5 Prior Sensitivity

As described in Section 1.2, we could use a variety of methods to choose a suitable prior to apply before undertaking Bayesian analysis procedures. In this project, we are focussing solely on vague priors, as we have no prior information about the parameters. One restriction with the WinBUGS program is that we cannot use any improper priors. Therefore we could not use, for example, a *Uniform*$[0, \infty)$ distribution as a prior; instead we must use distributions that naturally have support on $[0, \infty)$ and integrate to 1. So far, we have concentrated on using a vague *Gamma*(1, 0.0001) prior to analyse the data.

This is effectively an *Exponential*(0.0001) distribution but written in the Gamma form to exploit the conjugate nature of the Gamma distribution to the Poisson-Pareto model. Now we investigate the sensitivity of the Pareto model to the form of prior used. For more discussion about prior sensitivity in Bayesian inference, see [1].

We will compare the convergence and the resulting posterior distributions of the parameters using *Gamma*(0.001, 0.001), *Exponential*(0.001), *Uniform*(0, 10000), *Lognormal*(0, 1000) and *Normal*(0, 1000) (truncated to positive values only) as priors. For β, we restrict the range to [0, 1.625] as before. These distributions all have large variances, so are suitably vague, but the distributions themselves have different shapes. We do not consider a Pareto prior, as we do not wish to force a lower bound on the possible parameter values. The posterior distributions for α, β, θ and the average claim are illustrated in Figure 2.5.1.

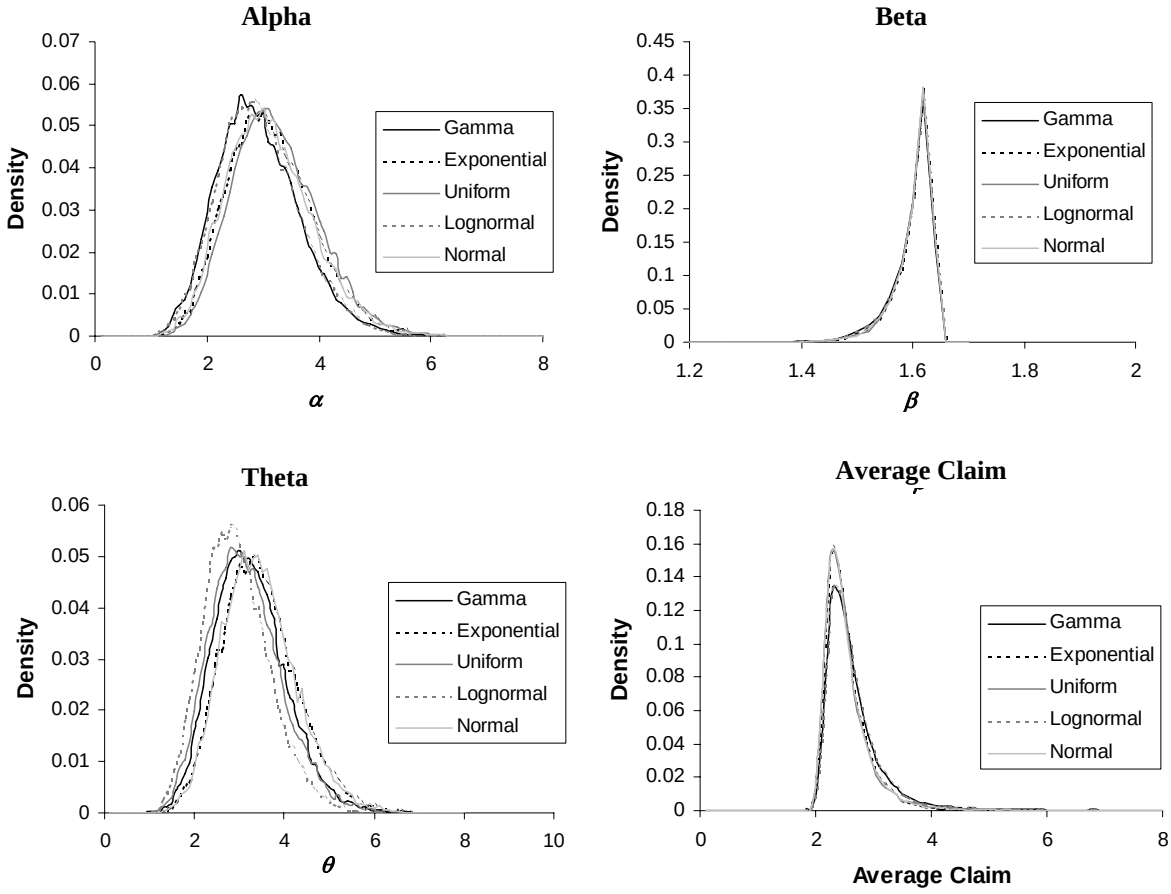

Figure 2.5.1 Posterior distributions of the parameters in the Poisson-Pareto model using different forms of prior distributions. Source: own calculations.

As we can see, these different priors all give very similar distributions, suggesting that the model is not very sensitive to changes in the prior distribution. We can separate the distributions into two groups. The Exponential, Uniform and Normal priors have posterior means of approximately 3.08 for α, 1.591 for β and 2.50 for the average claim (which is very similar to the results in Table 2.2.2), whereas the Gamma and Lognormal priors have 2.88 for α, 1.588 for β and 2.64 for the average claim. The posterior variances for all the parameters are very similar. It is quite interesting that if we compare the posterior means for the parameter θ, the distributions divide into two different groups. The Gamma, Uniform and Lognormal priors have the posterior mean for θ clustered around 3.2 (the sample mean), and the other two prior distributions have posterior mean close to 3.4 (the value given by the model we used in Section 2.2 and also the posterior mean from equation 7). There does not appear to be a pattern for either split. For example, the Gamma and Exponential distributions give different values for all the parameters, yet are closely related and have the same prior variance of 1000, so are equally "vague".

Comparing the rate of convergence of the chains using the different starting distributions is also quite informative. We estimate the burn-in period from the BGR diagram, and look for the first iteration for which the between and within chain variations are stable in value. This is an approximate method, but gives us a good indication of the speed of convergence. Out of the priors we have used here, the fastest to converge was the Exponential, closely followed by the Normal. The Lognormal was the slowest to converge, followed by the Uniform. There does not appear to be any relationship between the speed of convergence and the converged values. The values listed above are all taken from the chains once they have converged, and are the means calculated from 60000 iterations.

To summarise, the Poisson-Pareto method seems to be robust against the distribution of the vague prior selected.

2.6 Comparing Classical and Bayesian results

We can also compare the posterior mean results with estimates derived from classical methods. Examples include the Method of Moments Estimate (MME), in which we equate the first k moments to estimate k parameters, and the Maximum Likelihood

Estimate (MLE), in which we calculate the values of the parameters that will maximise the likelihood of observing the data. We have calculated the MME and MLE analytically for the standard distributions mentioned in Sections 2.2 and 2.3 when possible. We do not include the Weibull distribution, as we can only estimate the parameters numerically, and we discount the truncated forms of the distributions for this section. The estimates are given in Table 2.6.1.

Table 2.6.1 MLEs and MMEs for Pareto, Gamma, Exponential and Lognormal distributions. Source: own calculations.

Claim size model	α		β	
	MLE	MME	MLE	MME
Pareto	3.076	2.225	1.625	1.683
Exponential	0.3271	0.3271	N/A	N/A
Gamma	-	0.5014	-	0.1640
Lognormal	0.8106	0.5691	0.9921	1.097

In addition, as mentioned in Section 2.2, the MLE (and MME) of the Poisson distribution is 3.2. Figures 2.6.1 and 2.6.2 compare the distributions with these parameters to those with parameters equal to the posterior means (listed in Tables 2.2.2 and 2.3.1).

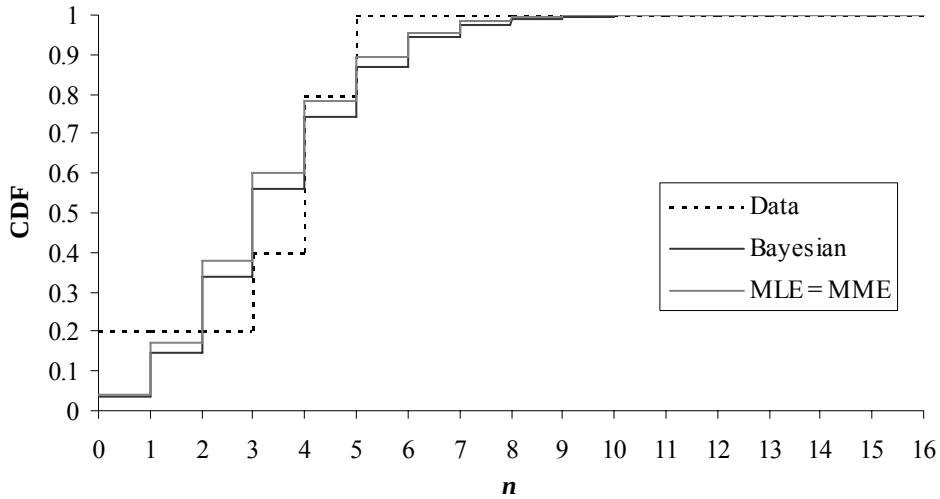

Figure 2.6.1 Comparison of fitted Poisson distributions to cumulated distribution of data n_t. Source: own calculations.

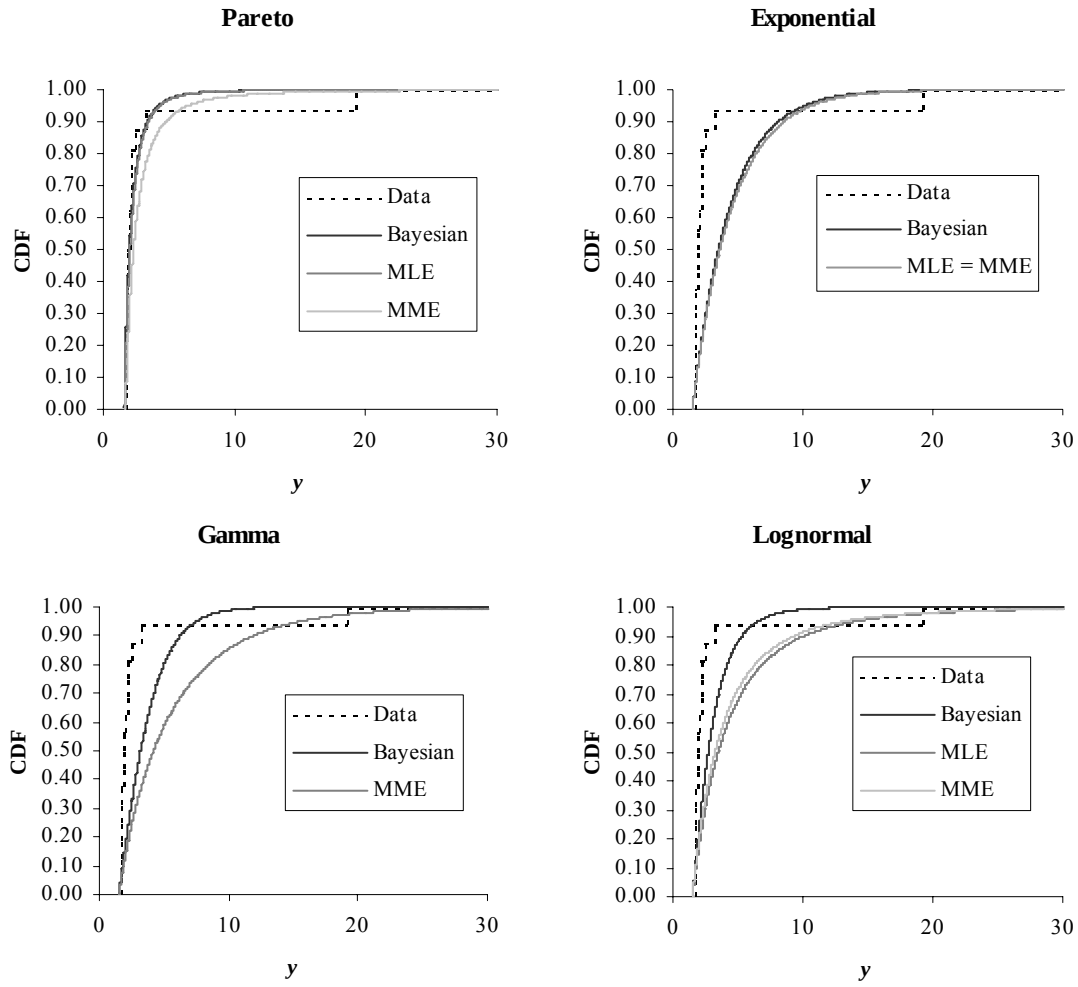

Figure 2.6.2 Cumulative density functions of various fitted loss distributions compared to cumulated distribution of data $y_{i,t}$.
Source: own calculations.

From examining these graphs, we can see that for these data, the Bayesian estimates from the MCMC method generally provide distributions with a better fit to the data than the simple classical methods. This is because the Bayesian approach takes account of more levels of variability in the model. Some of the classical estimates lie very close to the Bayesian results, such as the MLE for the Pareto distribution and the estimate for the Exponential distribution, but for these data, the Bayesian results are superior. The Bayesian methods also provide the whole posterior distributions for α, β and θ, which enable us to use predictive distributions and allow us to find more accurate approximations to N, Y and the future aggregate claim S. For more information about predictive distributions, see Section 3.

3 PREDICTIVE INFERENCE

In Chapter 2, we investigated some of the possible distributions we can use to model the data that we have collected for the aggregate risk model. In this chapter, we aim to obtain predictions of future aggregate claims S by combining the information for the number of claims, N, and the individual claim amount, Y. First, however, we shall examine the distributions of N and Y themselves. Throughout this section, we shall use the notation $\underline{\phi} = (\alpha, \beta, \theta)$ to denote the vector of parameters for our problem or any general vector of parameters, in order to keep results as general as possible.

3.1 Predictive Distributions Method

In order to learn about the distribution of S, we use the posterior predictive distribution. This takes into account the variability of $\underline{\phi}$ by integrating over all possible values of $\underline{\phi}$ using the posterior distribution as weights. It is a form of mixture distribution over the possible values of $\underline{\phi}$. This means that we use all the information provided by the Bayesian analysis and do not ignore the variation involved in $\underline{\phi}$, and so obtain better estimates than if we had merely used the posterior distribution. The posterior predictive distribution of some variable z (which is dependent on $\underline{\phi}$) given data $\underline{y}$ is defined by:

$$\pi(z \mid \underline{y}) = \int_{\Phi} f(z \mid \underline{\phi})\pi(\underline{\phi} \mid \underline{y})d\underline{\phi},$$

where $\pi(\underline{\phi} \mid \underline{y})$ is the posterior distribution of $\underline{\phi}$ given the data $\underline{y}$, and $f(z \mid \underline{\phi})$ is the density of the variable z given the value of $\underline{\phi}$. By using the predictive distribution, we ensure that we focus solely on observable values, rather than the unobserved parameter $\underline{\phi}$. We can interpret the predictive distribution as the density of a future observation.

Clearly, we cannot integrate directly over $\underline{\phi}$, as we do not have a closed form expression for the integrand, but instead we can use the output from the MCMC algorithm to obtain estimates of the posterior predictive distributions. See [9] for more information about the use of Monte Carlo integration in this context. We proceed as follows:
Let $\underline{N}$ be the set of data for N and let N_f be the value of a future observation of N. Then,

$$\begin{aligned}\pi(N_f = n \mid \underline{N}) &= \int_0^{\infty} p(N_f, \theta \mid \underline{N})d\theta \\ &= \int_0^{\infty} p(N_f \mid \theta, \underline{N})\pi(\theta \mid \underline{N})d\theta\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty p\left(N_f \mid \theta\right)\pi\left(\theta \mid \underline{N}\right)d\theta \\
&= E_{\theta|\underline{N}}\left\{p\left(N_f = n \mid \theta\right)\right\} \\
&= E_{\theta|\underline{N}}\left\{\frac{e^{-\theta}\theta^n}{n!}\right\} \qquad \text{as } N_f \sim Poisson(\theta) \\
&\approx \frac{1}{m}\sum_{i=1}^{m}\frac{e^{-\theta^{(i)}}\left(\theta^{(i)}\right)^n}{n!},
\end{aligned}
$$

where m is the number of MCMC simulations after convergence, and $\theta^{(i)}$ is the i^{th} simulated value of θ given $\underline{N}$.

Similarly, letting $\underline{Y}$ be the set of data for Y and y_f be the value of a future claim Y,

$$
\begin{aligned}
\Pr\left(Y \le y_f \mid \underline{Y}\right) &= \int_{\underline{u}} p\left(Y \le y_f, U = \underline{u} \mid \underline{Y}\right)d\underline{u} \qquad \text{where } \underline{u} = (\alpha, \beta) \\
&= E_{u|\underline{Y}}\left\{p\left(Y \le y_f \mid \alpha, \beta\right)\right\} \approx \frac{1}{m}\sum_{i=1}^{m}\left\{1 - \left(\frac{\beta^{(i)}}{y_f}\right)^{\alpha^{(i)}}\right\},
\end{aligned}
$$

where $\alpha^{(i)}$ and $\beta^{(i)}$ are the i^{th} simulated values of α and β respectively, given $\underline{Y}$.

These results give us a method of calculating approximate distributions for N_f and Y. The posterior predictive distribution for N_f is shown in Table 3.1.1.

Table 3.1.1 Posterior predictive distribution for future number of claims N_f, calculated from 30000 iterations. Source: own calculations.

n	0	1	2	3	4
$\Pr(N_f = n)$	0.0450	0.1277	0.1916	0.2023	0.1685
n	5	6	7	8	9
$\Pr(N_f = n)$	0.1180	0.0721	0.0394	0.0198	0.0091
n	10	11	12	13	14
$\Pr(N_f = n)$	0.0039	0.0017	0.0006	0.0002	0.0001

By comparing these figures to the values listed in Pai [15], we see that the probabilities are very similar. Using the predictive distribution rather than merely a Poisson distribution with parameter equal to the posterior estimate of θ means that we do not ignore the variability in the value of θ, so we obtain a more accurate distribution for N_f.

Since Y is a continuous random variable, we must discretise it to give us a set of values that we may use for simulation. We have discretised the sample space of Y to a set of

100 intervals, each with approximately equal probability. Figure 3.1.1 shows the posterior predictive density for Y using the Pareto model and the Truncated Weibull model. We have chosen to illustrate the Truncated Weibull model as comparison rather than the slightly better fitting Truncated Gamma model, due to ease of calculation.

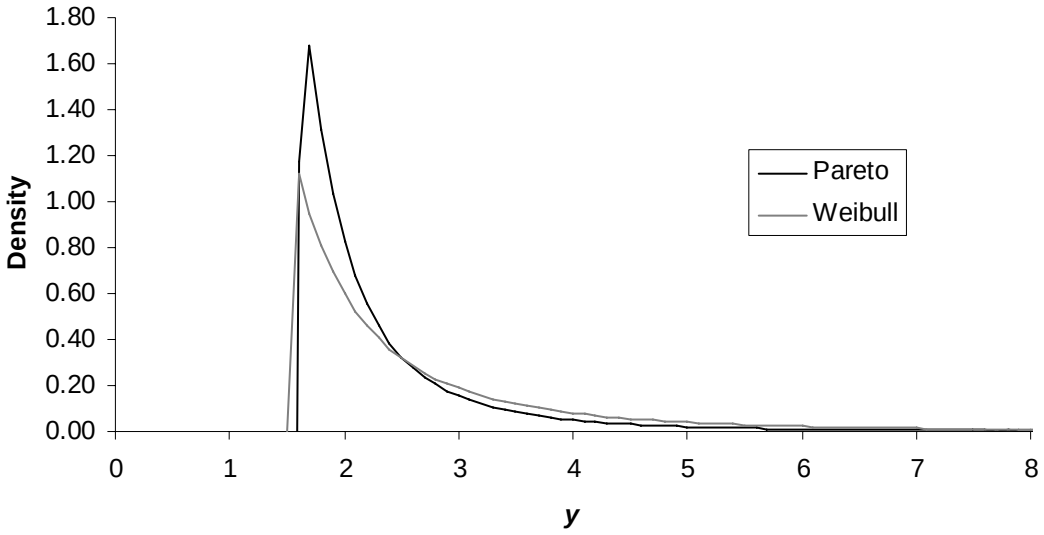

Figure 3.1.1 Posterior predictive density of individual claim amount Y for Pareto and Truncated Weibull models. Source: own calculations.

These two predictive models have similar shapes, but the Weibull model gives a fatter tail in this case than the Pareto model. We shall first examine the distribution of S using the Pareto model, then compare the results to the Weibull model.

3.2 Distribution of *S* using Pareto model for *Y*

We simulate a value for the future aggregate claim S as follows:

- Simulate a value n_f from the posterior predictive distribution for N_f.
- Simulate n_f values from the posterior predictive distribution for Y.
- Sum the n_f values of Y to obtain a value for S.

We sampled the distributions of N_f and Y by using the inverse cumulative probability function method. The values of S given by this method are approximate, as we are treating a continuous random variable, Y, as a discrete one. By using a larger set of values to discretise Y, we could obtain a more accurate distribution for S. Figure 3.2.1 gives a histogram of 10000 simulated values from the distribution of S using this method. We fit the overlaid distributions by equating the first two moments of the distribution to those of the simulated values. The Gamma distribution appears to fit the

best out of the three tried, although we cannot expect any standard distribution to model S precisely. The parameters of this distribution are $\alpha = 2.708$ and $\beta = 0.375$.

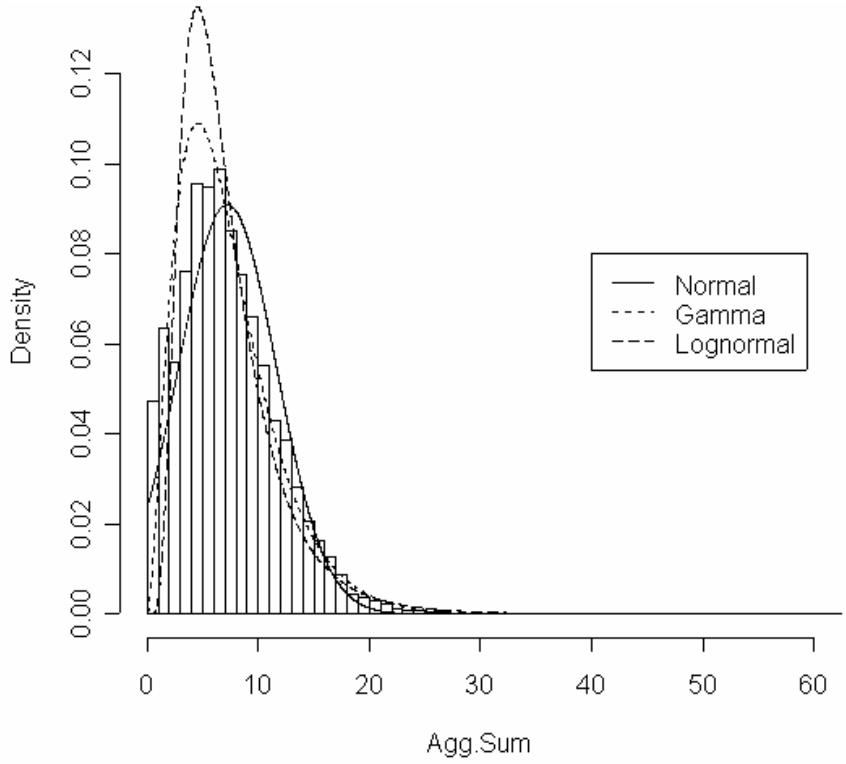

Figure 3.2.1 Histogram of simulated values of S using an approximation to the posterior predictive distribution, with possible distributions overlaid. Source: own calculations.

The mean and variance from these simulated values are 7.219 and 19.246, which are smaller than the sample mean and variance of S (9.784 and 84.129 respectively). We would expect the variance of the simulated values to be smaller than the sample variance, as one of the observed values of S is particularly large, and the Bayesian methodology has a "shrinkage" effect. That is, if there are any unusually large observed values, they are weighted appropriately so they do not dominate the analysis. The difference in the mean and variance are also partly because the discretised distribution for the individual claim amount Y is not particularly accurate in the tail, so very large values of Y do not appear in the simulations.

Table 3.2.1 shows the values of some percentiles from the simulated values. The maximum value appears to be relatively small, considering the claim size distribution should have a fat tail. This is due to the approximation used, which prevents large values of Y being simulated using this method.

Table 3.2.1 Estimated percentiles of approximate posterior predictive distribution of S. Source: own calculations.

Median	90th percentile	95th percentile	99th percentile	Maximum
6.61	13.09	15.32	19.51	33.64

We want to have a long tail to match the possible claim patterns that we may see – in most years, the aggregate claim will be relatively small, but occasionally a very large value will occur. The insurance company uses the information gathered about the frequency of these large claims and ensures that in the majority of circumstances it will have sufficient resources to pay for all claims. Otherwise, it will find itself in financial difficulties. For example, the company may decide to hold sufficient reserves to cover the aggregate claim for 99% of possible scenarios, in which case it would hold approximately 20 million (using these figures). If it holds too large a value in reserves, it loses out on the interest it could have earned on them. If it holds too small a value, it runs the risk of having insufficient reserves to fulfil its obligations. We therefore look at another method of simulating values from the predictive distribution, aiming to model the tail of the distribution of S more accurately.

3.3 Alternative Predictive Distributions Method using Pareto model for Y

In order to avoid the problems that occur by discretising a continuous distribution, we may progress in a different way. The procedure we use is as follows:

- Obtain estimates $\theta^{(i)}$, $\alpha^{(i)}$ and $\beta^{(i)}$ as before from the posterior distributions for θ, α and β.
- Simulate $N_f \sim Poisson(\theta^{(i)})$.
- Simulate N_f values $Y \sim Pareto(\alpha^{(i)}, \beta^{(i)})$.
- Sum these N_f values of Y to give $S^{(i)}$.

Here $\alpha^{(i)}$, $\beta^{(i)}$ and $\theta^{(i)}$ are the i^{th} simulated values of α, β and θ from the MCMC.

We repeat this for a large number of iterations, once the chains have converged to the equilibrium distribution. Collating the estimates $S^{(i)}$ gives us the same distribution asymptotically as the method in Section 3.2, but avoids the error involved in discretising the distribution of Y. The histogram of simulated values of S, together with a few possible distributions to model S, is shown in Figure 3.3.1. The mean and variance of

the simulated values of S in this case are 8.426 and 50.058. These values are larger than the mean and variance obtained from the previous method, which is due to the possibility of obtaining some very large claims, which was not the case before. This models the situation more accurately than the method of the previous section in which we had to impose an upper limit on the individual claim amount Y. This time, none of the overlaid distributions appear to fit the simulations particularly well.

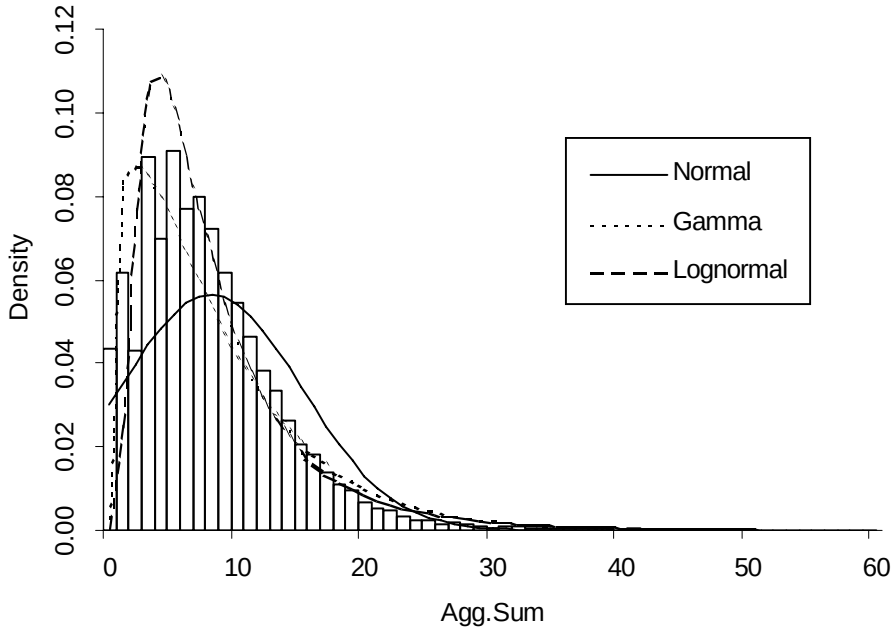

Figure 3.3.1 Histogram of simulated values of S using samples from the individual posterior distributions as parameter values, with possible distributions overlaid. Source: own calculations.

A few features of the posterior distribution of S are not immediately obvious from examining Figure 3.3.1. Firstly, there is a probability mass at zero, so the distribution of the future aggregate claim S has a mixed form, with a discrete point at zero and a continuous density from 1.5 to ∞. This is obvious from contemplating the problem posed. The probability at zero is equal to the probability of obtaining no claims in a year. Secondly, there are local minima in the density curve near multiples of 1.5. This is because individual claims must be of amount greater than 1.5 in this model and thus some small values of the aggregate claim are naturally more probable than other similar values. This effect is most noticeable near $S = 3$ in the histogram above.

Table 3.3.1 shows the values of some percentiles from this set of simulated values of S. Comparing to the values from Table 3.2.1, we see that the values are stretched over a wider range. This reflects the ability of this second method to model values in the tail

of the distribution more accurately. However, the 90th and 95th percentiles are of a similar order of magnitude to before, which shows that the two methods do give similar results for moderate values of *S*.

Table 3.3.1 Estimated percentiles of posterior predictive distribution of *S*. Source: own calculations.

Median	90th percentile	95th percentile	99th percentile	Maximum
7.33	15.49	18.84	28.19	228.69

3.4 Distribution of *S* using Truncated Weibull model for *Y*

For comparison purposes, we use the same method as in Section 3.3 to simulate the distribution of *S* with the Truncated Weibull model for *Y* rather than the Pareto model. Figure 3.4.1 plots the histogram of *S* for this model. The mean and variance of the simulated values are 9.67 and 59.47, which are similar to the Pareto model, but slightly larger. This may be explained from the fact that the Truncated Weibull model has a fatter tail than the Pareto model. The shape of the distribution of *S* is similar to that in Figure 3.3.1, although it is slightly more spread out as, for example, the 99th percentile is 34.84 compared to 28.19. This means that the insurers would hold slightly larger reserves if they used the Truncated Weibull model rather than the Pareto.

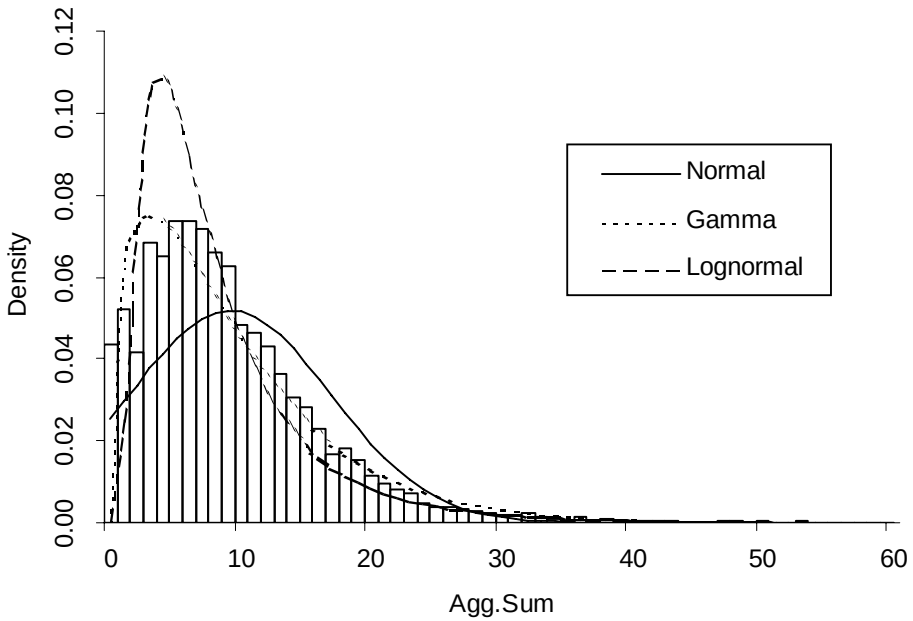

Figure 3.4.1 Histogram of simulated values of *S* using the Truncated Weibull model for *Y*, with possible distributions overlaid. Source: own calculations.

4 CONCLUSIONS

In this dissertation we set out to show how Bayesian methodology, combined with the use of MCMC methods, may be applied to the problem of forecasting future claims. We aimed to find a suitable distribution to model a particular instance of aggregate claims in general insurance. This distribution could then be used by insurers in predicting future aggregate claims, which would enable them to set suitable values for premiums and reserves. The methods used in this project are general, and we may implement them for any similar set of data.

In the first chapter, we presented a general background to Bayesian statistics, and set out the computational methods that we would apply later in the project, in particular Gibbs sampling through use of the WinBUGS package.

In Chapter 2, we fitted an aggregate claim model to data from a motor portfolio, assuming a Poisson distribution for the number of claims and employing various loss distributions for the single claim amounts. Our analysis suggested that the Poisson-Pareto aggregate risk model had the best fit to the data observed. Actuaries often use the Pareto distribution to model individual claim sizes, particularly in the case where there is a form of franchise deductible, as occurs here. Our results also showed that, for these data, the Pareto distribution is relatively robust to various assumptions of vague priors used in the Bayesian analysis. We then compared the findings from this analysis to those derived by classical methods, and found that the Bayesian results were superior in fit.

In the third chapter, we used the MCMC output to derive the posterior predictive distribution of the future aggregate claim, S, in two different ways that gave very similar results. As expected, the technique avoiding an approximation to the posterior distribution of the individual claim amount, Y, (detailed in Section 3.3) provided better results. We found that due to the nature of the deductible element in the model (which gives rise to local minima in the density) and the point mass at zero, we cannot find a simple distribution to model S, although it appears to be broadly similar in shape to a loss distribution, such as Gamma or Lognormal. Finding an analytical form for the distribution of S, or even for just the tail of S, would simplify the task of acquiring further information that may be of use to the

insurer. However, the techniques used in this project provide a method to obtain a very close numerical approximation to the distribution of *S*, obviating the need to fit an analytical density.

There are many areas we could investigate in more detail using these data. For example, we could try other loss distributions, such as the Burr, to model the individual claim amount, *Y*. Using distributions with more parameters might enable a better fit to the data. We may also test the fit formally by use of Bayes factors or deviances, rather than comparing distributions graphically. In addition, we could consider alternative distributions for the number of claims, *N*, and investigate how this affects the distribution of the future aggregate claim, *S*.

APPENDIX 1 - Probability density functions of distributions used

The truncated distributions are all truncated at $y = 1.5$.

Pareto(α, β):

$$f(y \mid \alpha, \beta) = \alpha \beta^{\alpha} y^{-(\alpha+1)}, \qquad y > \beta,\ \alpha > 0,\ 0 < \beta < \min\{y_{i,t}\}.$$

Gamma(α, β):

$$f(y \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y), \qquad y > 0,\ \alpha > 0,\ \beta > 0.$$

Truncated Gamma(α, β):

$$f(y \mid \alpha, \beta) = \frac{\beta^{\alpha} y^{\alpha-1} \exp(-\beta y)}{\Gamma(\alpha)(1 - \Gamma(\alpha; 1.5\beta))}, \qquad y > 1.5,\ \alpha > 0,\ \beta > 0,\ \Gamma(\alpha; x) = \int_0^x t^{\alpha-1} e^{-t}\, dt.$$

Exponential(α):

$$f(y \mid \alpha) = \alpha \exp(-\alpha y), \qquad y > 0,\ \alpha > 0.$$

Truncated Exponential(α):

$$f(y \mid \alpha) = \alpha \exp\{-\alpha(y - 1.5)\}, \qquad y > 1.5,\ \alpha > 0.$$

Lognormal(α, β):

$$f(y \mid \alpha, \beta) = \frac{1}{y\sqrt{2\pi\beta}} \exp\left\{-\frac{(\log y - \alpha)^2}{2\beta}\right\}, \quad y > 0,\ \alpha > 0,\ \beta > 0.$$

Weibull(α, β):

$$f(y \mid \alpha, \beta) = \frac{\beta}{\alpha} y^{\beta-1} \exp\left(-\frac{y^{\beta}}{\alpha}\right), \qquad y > 0,\ \alpha > 0,\ \beta > 0.$$

Truncated Weibull(α, β):

$$f(y \mid \alpha, \beta) = \frac{\beta}{\alpha} y^{\beta-1} \exp\left\{-\frac{1}{\alpha}\left(y^{\beta} - 1.5^{\beta}\right)\right\}, \quad y > 1.5,\ \alpha > 0,\ \beta > 0.$$

APPENDIX 2 – Computer code for WinBUGS for the Poisson-Pareto Model

```
model
{
      for(i in 1:16) {
                  y[i] ~ dpar(alpha, beta)
                  }
      for(i in 1:5) {
                  n[i] ~ dpois(theta)
                  }
      alpha ~ dgamma(1, 0.0001)
      beta ~ dgamma(1, 0.0001)I(, 1.625)
      theta ~ dgamma(1, 0.0001)
}
list(                          # Load data
      y = c(2.495, 2.120, 2.095, 1.700, 1.650, 1.985, 1.810, 1.625,
3.215, 2.105, 1.765, 1.715, 19.180, 1.915, 1.790, 1.755),
      n = c(5, 3, 4, 0, 4)
)
list(                          # Set initial values for a chain
      alpha = 3.076,
      beta = 1.62499,
      theta = 3.2
}
```

APPENDIX 3 – Computer code for WinBUGS for the Truncated Gamma model

```
model
{
    for(i in 1:16) {
        zeros[i] <- 0
        zeros[i] ~ dpois(phi[i])
        phi[i] <- -alpha*log(beta) - (alpha - 1)*log(y[i]) +
            beta*y[i] + loggam(alpha) + log(1-gam[i,
            21]/exp(loggam(alpha)))
        gam[i, 1] <- 0
# Specify the Incomplete Gamma function gam[i, 21] (only use the
first 20 summands from the infinite sum as approximation)
        for(j in 2:21) {
            gam[i, j] <- gam[i, j - 1] + exp(-1.5*beta) *
                pow(1.5*beta, alpha) * exp(loggam(alpha)) *
                pow(1.5*beta, j-2) / exp(loggam(alpha + j-1))
            }
        }
    alpha ~ dgamma(1, 0.0001)
    beta ~ dgamma(1, 0.0001)
}
```

The zeros method (used here) allows us to use arbitrary sampling distributions with log likelihood $l(i)$, by setting our observed data to be a series of zeros, and using the result that the likelihood of a $Poisson(\phi)$ observation of zero is $e^{-\phi}$. Setting $\phi(i) = - l(i)$ gives us the correct contribution to the likelihood. We can add constants to $\phi(i)$ without altering the log likelihood, to ensure that $\phi(i) > 0$ (which must hold, as it is a Poisson mean).

REFERENCES

[1] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.

[2] Casella, G. & Berger, R.L. (1990). *Statistical Inference*. Wadsworth and Brooks/Cole, Pacific Grove, California.

[3] Bayes, T. (1958). An essay towards solving a problem in the doctrine of chances. *Biometrika*, **45**, 293-315. (Reprinted from *Phil. Trans. Roy. Soc.,* **53** (1763), 370-418, with biographical note by G.A. Barnard.)

[4] Laplace, P.S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

[5] Jeffreys, H. (1939). *Theory of Probability*. OUP.

[6] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. A*, **231**, 289-337.

[7] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machine. *J. Chem. Phys.*, **21**, 1087-1091.

[8] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, **6**, 721-741.

[9] Gilks, W.R., Richardson, S. & Spiegelhalter, D.J., eds. (1996*)*. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

[10] Makov, U.E., Smith, A.F.M. & Liu, Y.-H. (1996). Bayesian methods in actuarial science. *The Statistician* **45** (4), 503-515.

[11] Rytgaard, M. (1990). Estimation in the Pareto distribution. *ASTIN Bulletin* **20** (2), 201-216.

[12] Panjer, H.H. & Willmot, G.E. (1983). Compound Poisson Models in Actuarial Risk Theory. *Journal of Econometrics* **23**, 63-76.

[13] Hogg, R.V. & Klugman, S.A. (1984). *Loss Distributions*. Wiley, Toronto.

[14] http://www.mrc-bsu.cam.ac.uk/bugs

[15] Pai, J.S. (1997). Bayesian Analysis of Compound Loss Distributions. *Journal of Econometrics* **79**, 129-146.

[16] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-472.

[17] http://www.mrc-bsu.cam.ac.uk/bugs/classic/coda04/readme.shtml

[18] Barnett, V. (1973). *Comparative Statistical Inference*. Wiley.

[19] Gibson, G.J. Notes on Bayesian Statistics. Heriot-Watt University.

ACKNOWLEDGEMENTS

To my parents, for their unceasing support throughout my time at university.

With thanks to Dr. George Streftaris and Prof. Gavin Gibson for their help and availability during this project, and for teaching me more about this fascinating subject area.