An overview of approaches to insurance data analysis and suggestions for warranty data analysis

Ming Luo, Shaomin Wu[1]

Kent Business School, University of Kent, Canterbury CT2 7PE, UK

Abstract

Warranty shares similarities with insurance in many aspects. Research on insurance data analysis has attracted much more attention than on warranty data analysis. This paper provides a general comparison between warranty and insurance in terms of their coverages, policies and data collection. It then reviews existing approaches to insurance data analysis with regard to modelling of claim frequency, modelling of claim size and policy pricing. Some recent patents relating statistical models are also discussed. The paper concludes with suggestions for improving warranty data analysis.

Key words: insurance claim, warranty claim, statistical modelling

1. Introduction

Warranty is a contractual obligation incurred by a manufacturer (vendor or seller) in connection with the sale of a product [1]. It is used for establishing liability in the event of an item's premature failure or the item's inability to perform its intended function [2]. A typical warranty transaction is: a consumer pays the warranty price when purchasing an item (items); then the manufacturer provides free repair or replacement service for the failures of the item(s) that occur during the warranty period.

Warranty can also be seen as a guarantee or promise about the reliability of a product. Warranty expense belongs to the operational expenses of the warranty provider who needs to cover labour and parts costs for repairs within the warranty period. Hence, unanticipated failures of sold items within warranty coverage may cause losses, which may include economic losses and reputation damage, to the warranty provider. Therefore, accurately forecasting the expected liability due to warranty claims is required in warranty data analysis, which involves modelling of the claim frequency and severity [3].

[1] Corresponding author. Email: s.m.wu@kent.ac.uk. Telephone: 0044 1227 827940
Accepted by journal *Recent Patents on Engineering* (in press).
DOI: 10.2174/1872212110666160617092705

Insurance is a tool through which the risk of a loss can be transferred from an encountered entity to another in exchange for payment. The principle of insurance can be explained as an insurer raising funds from a group of similar policyholders to pay the policy-holders who suffer losses within the group. The money raised from the policyholders is called premium, and the money paid for the claims is called claim amount/size/severity. An insurance company should hold the balance between the total premium and the total claim amount. The expected total claim amount is always treated as the expected liability of insurance companies. Insurance companies normally wish to forecast the expected total claim amount more accurately [4].

Warranty is a type of insurance. Hence, warranty and insurance data analyses share similarities. For example, the compound Poisson process is used in both warranty data analysis and the insurance data analysis. Nevertheless, there are differences between insurance and warranty in some aspects in which modelling approaches for analysing insurance claim data may not be applicable in warranty claim analysis. It has also been noticed that insurance modelling has attracted more researchers and is better studied than warranty modelling. As an evidence, on 15th January 2016, we searched the key words *warranty claim modelling* and *insurance claim modelling* in the "Article title, Abstract, Keywords" cell in www.scopus.com, and found 40 and 385 results, respectively. As such, this article will review state-of-the-art modelling techniques for insurance claims and learn what modelling techniques in warranty data analysis can be applied to warranty data analysis.

The rest of this paper is structured as follows. Section 2 compares the definitions, types and policies of warranty and insurance, respectively. Section 3 reviews the state-of-the-art modelling techniques in insurance data analysis. Section 4 suggests what modelling techniques from insurance data analysis can be learnt for improving warranty data analysis.

2. A general comparison between warranty and insurance

Although warranty is a type of insurance, they are different in some aspects. Below lists some of the differences that may need considering in data analysis.

2.1 Coverages

Warranty can be classified as two types, base warranty and extended warranty. Base warranty is provided when a product is sold, whereas extended warranty is purchased separately and voluntarily by the buyer as a service contract. That is, base warranty covers the early period of a product, starting from the first day when an item is put in use, whereas extended warranty covers a certain period after the base warranty has expired. Warranty can

be one-dimensional or two-dimensional. A one-dimensional warranty is characterized by an interval which is defined by time or usage, and a two-dimensional warranty is characterized by a region in a two-dimensional plane which consists of the time axis and the usage axis [5].

Insurance can be classified as life insurance and non-life insurance by the perils they insured. The peril insured by life insurance is death, and non-life insurance typically protects the insured from loss or damage caused by specific risks. Non-life insurance is a concept used in continental Europe which covers all insurance products except life insurance, but the same concepts used in the UK and the US are general insurance and property and casualty insurance respectively [5]. Insurance covers the period that both the insurer and the insured agree on.

2.*2 Policies*

A warranty policy is characterized by both duration and scope (full/limited, labour/parts, which parts, replacement/repair, etc.). There are four commonly used types of warranty policies for consumer goods: renewing free-replacement warranty, renewing pro-rata warranty, non-renewing free-replacement warranty, and non-renewing pro-rata warranty. Under a renewing warranty policy, an item that fails within its warranty period is repaired and the warranty is also renewed. Under a non-renewing warranty, the original warranty is not altered by a failed item and the warranty provider only guarantees satisfactory service on the item within the original warranty period. Under a free replacement policy, the warranty provider agrees to provide free repair services within the warranty period. However, if an item that fails during its warranty period is repaired at a certain cost to the consumer, the policy is a pro-rata warranty policy.

An insurance policy is a legal document which confers the policyholder the right to make claims and obliges the policy issuer to accept claims when covered events occur [4]. In life insurance, during the coverage period of one policy, the number of claims may not be greater than one (i.e., one can only die once). However, in warranty policy or non-life insurance, it is common that the number of claims is more than once during the coverage period. Hence, in this review, comparison is conducted between warranty and general insurance.

2.3 Data

Warranty data consist of claims data and supplementary data. Warranty claims data are lifetime data and collected during the servicing of warranty claims [7], which commonly include manufacture date, manufacture volume, sales date, sales volume, claims date and claims cost. Warranty claims data are the data collected when warranty claims happen. The supplementary

data are additional data, which are not always collected. Such data may be from those unfailed items whose warranty has not been claimed, even though such data may be analysed [3].

General insurance data are normally recorded by the insurer as a two-way array includes cases and variables. In general insurance, a claim is a demand for payment of damages covered under an insurance policy [8]. The claim size is the severity of the effect associated with claims and may be referred to in the literature as severity, loss, size or amount of damage and cost of a claim. The policies, claims, policyholders, accidents, etc. can be denoted as cases, while the level of injury, gender, claim amount, etc. can be recorded as variables [9].

In insurance, variables in claim reports can be quantitative or qualitative. For example, in the datasets used by [9], the variables recorded in personal injury insurance include claim amount, injury type, accident date, reporting date and finalization date. Variables recorded 93 in one-year vehicle insurance may include policyholder's age and gender, area of residence, vehicle value, vehicle age, vehicle type, claim occurrence and claim amount. Sometimes, the original records of insurance policies contain more detailed information. For instance, in the study of [10], the data collected from an insurance company in Singapore include vehicle and driver characteristics, insurance coverage and annual claims experience. Insurers normally record detailed information about policies and claims for accounting and premium rating purposes. The original data are tedious, the distribution of claim size and other data can be estimated through data pre-processing. Moreover, the claims data are often limited as data may be missing or corrupted [11].

Compared with warranty data, insurance data may contain more information such as information of the policyholders and they are normally panel data [12, 13], whereas warranty data are normally not presented as panel data format. Insurance claims occur due to one of the specific perils agreed in the policy whereas warranty claims occur due to product failures. Insurance data provide longer term information than warranty data, for example, car owners may need to purchase car insurance annually, even after the car's warranty has expired.

As a summary, Table 1 presents the comparison between insurance and warranty claims data.

Table1

Comparison between insurance and warranty claims data

	Insurance	**Warranty**
Number of variables	many (including the characteristics of policyholders)	few (only including product and repair data)
Occurrence	usually once	maybe recurrently
Claim amount	maybe large	normally small
Causes	perils (accidents)	failures
Covered interval	may cover the whole life of a product	the early life of a product

Warranty claims data and insurance claims data may contain basic information of claims such as claim size and occurred date. However, insurance claims data may have more variables/covariates than warranty claims data, for example, insurance data may include policy holders' information such as the driver's gender, age and other characteristics whereas warranty data do not contain such detailed information.

It should be noted that the quality of warranty data, which are normally collected from field, may not always be perfect, as they may be aggregated, delayed and censored.

3. Modelling techniques for insurance claims

Insurance data analysis focuses on reserving modelling and premium rating. Reserving modelling aims to forecast claims reserve, which is an insurer's future obligation and equals to the present value of an insurance policy's future cash flows. The insurer's total liability is the sum of the claims reserves of all individual policies issued. Premium rating is a process within which the amount that policyholder should pay the insurer for an insurance policy is determined. In this process, various properties of the insured object and the policyholder are taken into consideration.

Reserving modelling in both warranty and general insurance data analyses involves three components, modelling of claim frequency, modelling of claim size or severity and modelling of total claim amount. The claim frequency is the number of claims during a given period; the claim size or severity is the monetary amount of each claim and the total claim amount is the sum of the monetary losses of all claims.

3.1 Modelling of claim frequency

In general insurance data analysis, the Poisson distribution and the Poisson process are benchmark models in modelling claim counts data, and the variants of those models are widely used [14]. For example, when the inter-arrival times between claims are independent and identically distributed according to the exponential distribution, with parameter λ, then the probability that there are k claims is given by

$$P\{N(t) = k\} = \frac{(\lambda \mathrm{T})^k e^{-\lambda \mathrm{T}}}{k!}$$

where $N(T) =$ cumulative number of failures from time 0 to time T.

In order to improve the performance of the modelling, one may also use other modelling techniques such as those reviewed below.

In [15], a generalized linear mixed model (GLMM) is applied to analyse repeated insurance claim frequency data. A conditionally fixed effect vector is incorporated into the linear predictor to model the inherent correlation. The study is based on a motor insurance dataset. The authors argue that the Poisson regression model is not proper for modelling the dependence between the observations from the same policyholder and they then extend the Poisson regression to a random effects model, in which a random effect vector is used for modelling the shared effects of repeated observations for the same policyholders.

[16] discuss fixed and random effects models used on longitudinal data in insurance claim modelling, with a case study based on three consecutive year data of a motor's third party liability insurance portfolio in Belgium. The data include the annual number of claims and the characteristics of the insured, such as sex and age of drivers, power of the vehicle and the size of city. The number of claims per year is assumed to follow the Poisson distribution with specific parameter for each policyholder. The parameter of the Poisson distribution is treated as a linear combination of the observable characteristics multiplying a static factor. In the random effects model, the static factor is expressed as a random variable with unit mean, and in the fixed effects model, the static factor is expressed as an estimated parameter of each individual. The parameters are estimated through the maximum likelihood. The static factors (i.e. heterogeneity parameters) are modelled by the Poisson-Gamma model, the Poisson-log normal model and the Poisson-inverse Gaussian model, respectively. They conclude that in a short period or with a few observations, the random effects model is better than the fixed effects model. However, in the data of the motor's third party liability insurance portfolio, the heterogeneity is not

identically distributed across the insured, which can lead to inconsistency in the random effects model. Then they relax the assumption that the heterogeneity term is i.i.d., and use a regression on the individual heterogeneity terms to link the fixed effects model to the random effects model.

In [17], the Frank copula is used to jointly model the type of coverage and the number of accidents, in which the dependence parameter of the copula depicts the relationship between the frequency of accidents and the choice of coverage. Based on the copula model calibrated by one-year cross-sectional claims data collected from a major Singaporean automobile insurer, they find a significant positive coverage-risk relationship. This method is able to help derive the pure premium through demonstrating the effect of coverage choice on the incidence of accidents. Similarly, [18] use copulas to model the dependence between the class occupied by the insured and the claim frequency, and take the zero-excess phenomenon into account.

[12] investigate longitudinal data models of claim counts with excess-zeros and consider the dependence between claims from two successive periods. Copula is used to model the dependence. They discuss two models: The first one is a discrete margin copula model, which is applied to fit the time series data of longitudinal observations. The second one uses a copula to model the time-dependent unobservable random effects. In the first model, they use the Poisson distribution to model the number of claims with a mean claim frequency that incorporates the covariate information. Considering that the insured may not claim on the two successive periods, they use the zero-inflated Poisson model. In the second model, the residual heterogeneity in tariff cells is considered and modelled by a random effect, which presented as a random effect variable in the mean of the Poisson distribution with excess zeros.

[13] provide a copula-based method to model the number of claims within a longitudinal context. The authors aim to predict the number of claims in a subsequent period based on the data in the previous periods. In order to reduce the computational difficulties, the claim amount is converted to a continuous random variable by a technique called jittering and then the joint distribution of the number of claims in successive periods is modelled through copulas. Jittering is a method in which an independent continuous random variable is subtracted from each component of the multivariate discrete data, aiming to convert discrete variables to continuous ones.

[19] introduces a non-homogeneous Poisson cluster model to cope with the reducing property of payment processes. The Poisson cluster process is used to model the total number or amount of payments for the claims arriving in a year and being paid in the interval $[0, t](t \geq$

1), with different clusters. The paper aims to predict the claims occurring in a future interval $[t, t+s](t \geq 1, s > 0)$. The author considers a model with additive Levy processes, non-homogeneous Poisson clusters and non-homogeneous negative binomial clusters.

3.2 Modelling of claim size

In general insurance data analysis, the exponential-inverse Gaussian distribution, which has a shorter tail than the Pareto distribution and allows for incorporating covariates, is introduced in [20] to model claim size. In their model the claim size is assumed to follow the exponential distribution, whose parameter is $\theta_i t_i$, where θ_i follows the inverse Gaussian distribution and the logarithm of t_i has a linear relationship with covariates. This model can be treated as an exponential regression model with random effects, in which the random effects are related to the inverse Gaussian distribution.

[21] consider the possible dependence between the claim size and the waiting time for a claim in an insurance portfolio, which differs from the conventional assumption in which time between claims and claim sizes are independent. The authors assumed inter-claim time and its subsequent claim size to be dependent according to an arbitrary copula structure.

[10] propose a Bayesian hierarchical approach to estimating joint multivariate distributions that model claim amounts of various claim types. The hierarchical model is decomposed to three components related to the frequency, type, and severity of claims, respectively. In this research, initially, a negative binomial regression model is used to assess claim frequency. The main contribution of this research lies in their introduction of a multivariate claim distribution to handle long-tailed, correlated claims with covariates.

[22] also investigate the independent assumption between claim types, and then apply the Gaussian, Frank and Clayton copulas to model the dependence between claim types. This research is based on a motor insurance claim dataset from Malaysia.

In order to estimate the loss reserves for incurred but not reported (IBNR) claims through individual claim loss models instead of aggregated claim loss models, [23] use the semi-survival copula and the semi-competing risk copula to model the dependence between the event times with delays in the individual claim loss model. The performances of their proposed methods are evaluated through simulation. In this research, the assumption, that relationship between the event time and the delay is independent, is relaxed; and the dependence structure between the event time and the delay is characterized by a copula with the margins of the event times and the delays.

[4] discard a crucial assumption that the number of claims and the claim sizes are independent. A mixed copula approach is provided to model the dependence between the number of claims and the average claim size through the Gaussian copula. The number of claims in a group of characterized policyholders is modelled by Poisson regression, and the average claim size of the group is modelled through Gamma regression. They then construct the joint distribution of the two marginal distributions through the Gaussian copula. They find that this model performs well overall, but the extreme values are not presented very well. This may be caused by copula selection which is suggested for future study.

To estimate the size of claim reserving, [24] develop a type of hierarchical Bayesian-based claim models that considers claims payments and incurred losses information. Their method extends the claims reserving models of [25] and [26]. They also use a data-augmented mixture Copula paid incurred claims model.

3.3 Policy pricing

In general insurance analysis, a conventional assumption is that the claim frequency and the claim size are mutually independent [14]. However, in practice, researchers find that this assumption is too restrictive.

Premium rating, or ratemaking, is the core activity in general insurance policy pricing. The key principle in ratemaking is cost-based pricing of individual risks. The price charged to the policyholders for the coverage in an insurance policy is the estimated present value of the future costs incurred by the covered peril. In the pure premium approach, the price of an insurance policy is defined as the ratio of the predicted costs of all future claims against the insurance policy coverage. The price is in effect to the risk exposure and expenses [27]. Additionally, the property ratemaking is based on the distributions of claim frequency and loss, which is similar to estimating the total reserves [14]. That is, ratemaking aims to price insurance policy based on individual's characteristics and claim history, in which the policyholder's related risk level and expected claims will be estimated, and the price and coverage of the insurance policy will be determined based on the estimation.

[27] state that actuaries have to design a tariff structure to fairly allocate the burden of claims among policyholders in a competitive market. The policies are categorised into classes. The policyholders in the same class are charged the same premium. In practice, property and vehicle insurers use risk classification plans to create classes. The classification variables are priori variables as those values are determined before the policyholder starts to use their property. Premiums for motor liability coverage are often set based on territory the vehicle

garaged in, the use of the vehicle, and driver's individual characteristics such as age, gender, occupation, etc. A priori classification can be achieved with generalized regression models.

Currently, in motor insurance, a popular approach is using experience rating to link premium amounts to individual's past claims experience. In this approach, the insured drivers who are responsible for accidents are penalized by premium surcharges (or maluses), and the claim-free policyholders are rewarded by discounts (or bonuses). The systems applying this approach are called Bonus-Malus systems, experience rating, no-claim discounts, or merit rating [27]. [17] also state that in general insurance ratemaking is a classical actuarial problem. In principle, the distribution of insurance claims is the basis for determining pure premiums.

All in all, the ratemaking approaches in insurance can be divided into two categories; the first one determines the risk classification of the policyholders based on the policyholders' supplementary information (covariates) and historical data of the similar policy, and the premium will be computed according to the risk classification. The second one is the experience rating which determines the premium based on the policyholder's individual past claims experience. These two approaches can also be applied together when the first one determines the basement of the premium and the second one makes correction on the basement.

[28] introduce an optimal Bonus-Malus System (BMS) in automobile insurance which considers the frequency and severity of a policyholder's historical accidents simultaneously instead of the major BMS designed based on the frequency of accidents and disregarding the incurred severity. In their optimal BMS the frequency of accidents and the severity of accidents are assumed independent. They model the frequency component of optimal BMS with the negative binomial distribution, which is a gamma-Poisson mixture distribution, and model the severity component with the Pareto distribution, which is a conjugated distribution of exponential and inverse-gamma distributions.

[29] applies the bivariate Poisson regression model in priori ratemaking. Priori ratemaking is conducted based on the priori variables, which are used to segment the insurance policies portfolio in homogeneous classes. [29] relaxes the assumption that the number of claims of different types are independent and uses bivariate Poisson regression to model the number of claims for third-party liability and the number of claims for other guarantees. The result shows the independence assumption between claim types should be rejected.

According to [17], insurance claim data have the unique semi-continuous feature in the sense that a positive continuous component is associated with a significant fraction of zeros. In the

actuarial practice, two methods are widely used as standards: the Tweedie generalized linear model (GLM) and the frequency-severity model. The former incorporates a percentage of zeros into a continuous distribution through a compound Poisson process, where the resulting Tweedie distribution is featured with a mass probability at zero and thus is suitable for modelling the semi-continuous claim data. The latter decomposes the total claims amount into frequency and severity parts. An insurance policy usually provides multiple types of coverage. The unique contract is designed for each type such as deductibles or coverage limits, insurers must deal with the payment separately. Plus, the insurers can obtain additional insights by decomposing aggregated claims into different categories and investigating their joint behaviour [30].

[31] find the independent assumption of different types of claims is not realistic in automobile insurance. They point out that in the classical priori tariff system not all risk factors are identified: the tariff classes can be heterogeneous and the unobserved heterogeneity and serial dependence may lead to over dispersion. Hence they apply a multivariate Poisson regression model such as the zero-inflated model. As their model has computational difficulties, they introduce Bayesian inference based on Markov chain Monte Carlo (MCMC) method to solve it.

[18] discuss the Bonus-Malus system in posteriori ratemaking. As an innovation, they use a bivariate copula function to model the dependence between claim frequency and policyholder's risk class. The marginal distribution of the number of claims is modelled with Poisson regression. The distribution of the classes is represented through a matrix and the joint distribution is modelled by the Clayton copula.

[32] review statistical tools used in risk classification for ratemaking. They state that, in general insurance, a priori ratemaking is used to build the basis when a policyholder is new and insufficient information is available, and a posteriori ratemaking is used to correct and adjust the priori premium when the historical information about policyholder becomes available.

4. Recent related patents

Normally, insurance claims provide more data whereas warranty claims provide less data. This is because more detailed data about policyholders are available in insurance data collection whereas it is impossible to record detailed information about customers in warranty data collection. With technology progressing, however, more product operation data can be collected. For instance, the radio frequency (RF) devices and microelectro-mechanical systems (MEMS), as

well as the advances in wireless technologies, provide a promising infrastructure for gathering information about parameters of the physical world.

Some recent patents can help improve warranty data collection and analysis. For example, patent [33] invent a NFC tag with new structure to extend the coupling range of the tag with external reader, which can help enhance the communication of information between management systems and the products. Patent [34] relates to methods, devices and computer-readable media for acquiring statistical access models. This patent aims to detect an anomalous access event of a designated user according to the merged statistical access model of two user groups' access profiles. The mechanism of [34] may be applied in warranty management to detect the anomalous warranty claims. Patent [35] describes a computer-implemented method and system for processing data. It provides a method of feature selection for the statistical models efficiently and effectively, which can facilitate creation and application of statistical models in business analytics. The mechanism of this patent can help perform warranty data modelling.

Generally, insurance data analysis and warranty data analysis are mainly concerned with modelling approaches and algorithms, there are many patents relating to relating to those issues. For example, patent [36] develops an insurance program for entities in the mortgage industry that provides coverage for financial loss as a result of material inaccuracies in the financial information provided by or on behalf of the borrower, patent [37] provides a method and system of predicting a maintenance schedule and estimating a cost for warranty service of systems, for example, hardware systems, patent [38] develops a system and method for organizing and analysing field warranty data, and patent [39] develops a method and apparatus for calculating warranty cost relating to various aspects.

5. Current and Future Developments: suggestions for warranty data analysis

The aim of warranty data analysis is to extract useful information and help in decision making. Warranty data analysis serves five areas [3], early detection of reliability problems, and suggestion on design modification, field reliability estimation, claim/cost prediction, and claim/cost estimation. The first three areas are mainly concerned with the reliability problems, on which there is no counterparts in insurance. It should be noted that warranty claim estimation is for a hypothetical in finite population of items, of which those sold are considered a random sample, whereas in warranty claim prediction, the population of items that is eventually sold is finite [3]. A better estimation of future warranty claim/cost is also needed for determining the warranty reserves [5].

5.1 Data collection

The above review conforms that more data are available in insurance data analysis than in warranty data analysis: observations of covariates relating to insurance claims are collected and widely used in insurance data analysis while observations of covariates relating to warranty claims are not often available and used in warranty data analysis. The patents reviewed in Section 4 allow for manufacturers to install wireless sensors to monitor the operating condition, the usage and the deterioration status of their products, which are covariates affecting the reliability of the products. That is, to collect observations of the covariates relating warranty claims is becoming much easier. Those collected data can be analysed with modelling techniques borrowed from insurance data analysis.

5.2 Dependence modelling

In the development process of insurance claim modelling, a conventional assumption is the independence of claims [15] on the ground of the fact that the insurance claims in a group of policies are caused by independent accidents. However, many researchers find that, in practice, the independence assumption is too restrictive and may not hold. As such, dependence between claims in different periods claimed by a policyholder is considered, based on which many modelling approaches are developed, see [9, 12-15, 40], for example.

5.*2.1 Dependence in warranty data analysis*

According to [3], most literatures on warranty claim data analysis for products with one-dimensional warranty policies do not consider the dependences within claims data.

When analysing one-dimensional warranty data, existing approaches include estimating mixed distributions [41], fitting the Weibull distribution [42], considering sales delay [43], etc. Those approaches do not consider any dependence within basic claims data, such as the dependence between claims and failure modes, which may cause the loss of information and result in biased decision making.

When analysing two-dimensional warranty data, some types of dependence are considered, for example, [44], [45] and [46] discuss modelling the dependence of failures on age and mileage, which directly estimates a joint bivariate distribution depicting the dependence between age and usage [47-50], the dependence between failures and age/usage [51], and the dependence between recurrent events and usage [52].

Having reviewed the existing papers in warranty data analysis and insurance data analysis, we find that copula-based approaches are widely used in insurance data analysis, but barely applied in warranty data analysis.

Copula is a tool that models the dependence structures between random variables. It was first introduced by [53] and has attracted considerable attention in theoretical and application aspects in recent years. In insurance data analysis, copula-based approaches are used to model the dependence between different claim types [10, 54], between accident date and reported date [23], between policy coverage and number of claims [17, 18, 55], between claim counts in successive periods [12, 13], between claims in different business lines [56, 57], between number 398 of claims and average claim size [4], and between time-to-claim and claim size [8].

In warranty data analysis, however, there is only one article using copulas to model dependence [50], in which a new method of constructing asymmetric copulas is introduced, and the asymmetric copulas are used to capture the tail dependence between the pair of age and usage in two-dimensional reliability data.

In general, the existing methods of warranty data analysis lack a systematic research on the dependence between warranty claims of different products and within warranty claims of the same products. Learnt from insurance data analysis, warranty data analysis can be improved by applying copula-based modelling approaches.

5.3 Random effects

In the literature, there are two existing approaches dealing with random effects, which are frequentist and Bayesian methods. The former includes linear modelling and the latter includes hierarchical Bayesian modelling methods, for example. Copulas, a recently widely studied method, falls in the former category.

As shown in the above review, insurance data analysis considering random effects has been studied by many authors, see [12, 13, 15, 16, 20], for example. In warranty data analysis, the phenomenon of random effects is studied as well. For example, [58] use flexible non-homogeneous Poisson processes as forecasting methods for warranty claims, where the possible heterogeneity among the products is modelled through random effects; [51] introduce a mixed Poisson process to model repeated events with both age and usage scales, and the heterogeneity in both usage and event rates across product units is accommodated with independently and identically distributed random effects. [59] assume that random effects exist

in warranty data due to the fact that products are manufactured by different production lines and warehoused in different regions with different environments.

One may also use copulas to analyse warranty data with a consideration of random effect, which can avoid information loss. As literature review conducted above, in insurance data analysis, the dependence of the claims frequency of heterogeneous units in successive periods is depicted by the dependence between the random effect variables and modelled by copula [12]. Hence there is a need to consider the random effects within unobserved heterogeneity in warranty data to avoid information loss.

5.4 Policy pricing

Warranty policies are extensively studied in the literature, but most of them are based on assumptions such as \warranted items being repaired minimally" or “being repaired minimally", which are very much concepts widely used in reliability and maintenance engineering, see [60,61], for example. As mentioned above, different from base warranty, extended warranty is purchased separately and voluntarily by the buyer as a service contract. This suggests that the price of the extended warranty may be optimised on the basis of operating condition, usage intensity, consumers' information, etc., which may be collected from a variety of sources in the era of big data. Based on such information, warranty providers will be able to sale extended warranties with different prices to different consumers. For this reason, modelling techniques of risk classification in policy pricing in insurance may be adapted to warranty data analysis.

6. Remark

This review has tried to be reasonably complete. However, those papers that are not included were either considered not to bear directly on the topic of the review or inadvertently overlooked. Our apologies are extended to both the researchers and readers if any relevant papers have been omitted.

7. Conclusions

This paper compares insurance and warranty analysis from different aspects, including comparison of modelling techniques and the development of patents relating to those analyses. It finds that copulas are becoming more widely used in insurance data analysis than in warranty data analysis. The paper then suggests some aspects in which copulas can be utilised in improving warranty data analysis.

8. Conflict of Interest Statement

The second author's time was partly supported by the Economic and Social Research Council of the United Kingdom (Project Ref: ES/ L011859/1).

9. Acknowledgement

We are grateful to the reviewers for their helpful comments, with which the clarity of this paper is improved.

Reference

[1] W. Blischke, Warranty cost analysis, CRC Press, 1993.

[2] W. R. Blischke, D. Murthy, Product warranty management: A taxonomy for warranty policies, European Journal of Operational Research 62 (2) (1992) 127-148.

[3] S. Wu, Warranty data analysis: a review, Quality and Reliability Engineering International 28 (8) (2012) 795-805.

[4] C. Czado, R. Kastenmeier, E. C. Brechmann, A. Min, A mixed copula model for insurance claims and claim sizes, Scandinavian Actuarial Journal 2012 (4) (2012) 278-305.

[5] W. R. Blischke, M. R. Karim, D. P. Murthy, Warranty data collection and analysis, Springer Science & Business Media, London, 2011.

[6] M. V. Wuthrich, M. Merz, Stochastic claims reserving methods in insurance, Vol. 435, John Wiley & Sons, 2008.

[7] S. Wu, A review on coarse warranty data and analysis, Reliability Engineering & System Safety 114 (2013) 1-11.

[8] P. Weke, C. Ratemo, Estimating IBNR claims reserves for general insurance using Archimedean copulas, Applied Mathematical Sciences 7 (25) (2013) 1223-1237.

[9] P. De Jong, G. Z. Heller, et al., Generalized linear models for insurance data, Vol. 136, Cambridge University Press, Cambridge, 2008.

[10] E. W. Frees, E. A. Valdez, Hierarchical insurance claims modelling, Journal of the American Statistical Association 103 (484) (2008) 1457-1469.

[11] P. Cizek, W. K. Hardle, R. Weron, Statistical tools for finance and insurance, Springer Science & Business Media, Heidellberg, 2005.

[12] X. Zhao, X. Zhou, Copula models for insurance claim numbers with excess zeros and time dependence, Insurance: Mathematics and Economics 50 (1) (2012) 191-199.

[13] P. Shi, E. A. Valdez, Longitudinal modelling of insurance claim counts using jitters, Scandinavian Actuarial Journal 2014 (2) (2014) 159-179.

[14] T. Mikosch, Non-life insurance mathematics: an introduction with the Poisson process, Springer Science & Business Media, Heidelberg, 2009.

[15] K. Yau, K. Yip, H. Yuen, Modelling repeated insurance claim frequency data using the generalized linear mixed model, Journal of Applied Statistics 30 (8) (2003) 857-865.

[16] J. Boucher, M. Denuit, et al., Fixed versus random effects in Poisson regression models for claim counts: A case study with motor insurance, Astin Bulletin 36 (1) (2006) 285-301.

[17] P. Shi, E. A. Valdez, A copula approach to test asymmetric information with applications to predictive modelling, Insurance: Mathematics and Economics 49 (2) (2011) 226-239.

[18] X. Zhao, X. Zhou, Copula-based dependence between frequency and class in car insurance with excess zeros, Operations Research Letters 42 (4) (2014) 273-277.

[19] M. Matsui, Prediction in a non-homogeneous Poisson cluster model, Insurance: Mathematics and Economics 55 (2014) 10-17.

[20] N. Frangos, D. Karlis, Modelling losses using an exponential-inverse Gaussian distribution, Insurance: Mathematics and Economics 35 (1) (2004) 53-67.

[21] H. Albrecher, J. L. Teugels, Exponential behavior in the presence of dependence in risk theory, Journal of Applied Probability 43 (1) (2006) 257-273.

[22] Y. Resti, N. Ismail, S. H. Jaaman, Handling the dependence of claim severities with copula models, Journal of Mathematics and Statistics 6 (2) (2010) 136-142.

[23] X. Zhao, X. Zhou, Applying copula models to individual claim loss reserving methods, Insurance: Mathematics and Economics 46 (2) (2010) 290-299.

[24] G.W. Peters, A. X. Dong, R. Kohn, A copula based Bayesian approach for paid-incurred claims models for non-life insurance reserving, Insurance: Mathematics and Economics 59 (2014) 258-278.

[25] J. Hertig, A statistical approach to IBNR-reserves in marine reinsurance, Astin Bulletin 15 (02) (1985) 171-183.

[26] D. Gogol, Using expected loss ratios in reserving, Insurance: Mathematics and Economics 12 (3) (1993) 297-299.

[27] M. Denuit, X. Mar510 echal, S. Pitrebois, J.-F. Walhin, Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems, John Wiley & Sons, Chichester, 2007.

[28] N. E. Frangos, S. D. Vrontos, Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance, Astin Bulletin 31 (01) (2001) 1-22.

[29] L. B. i Morata, A priori ratemaking using bivariate Poisson regression models, Insurance: Mathematics and Economics 44 (1) (2009) 135-141.

[30] P. Shi, Insurance ratemaking using a copula-based multivariate tweedie model, Scandinavian Actuarial Journal 3 (2016) 198-215.

[31] L. Bermudez, D. Karlis, Bayesian multivariate Poisson models for insurance ratemaking, Insurance: Mathematics and Economics 48 (2) (2011) 226-236.

[32] K. Antonio, E. A. Valdez, Statistical concepts of a priori and a posteriori risk classification in insurance, Advances in Statistical Analysis 96 (2) (2012) 187-224.

[33] C. MAK, C. LEUNG, Near field communication (nfc) tag, WO Patent App. PCT/CN2014/087,930 (Apr. 7 2016). URL http://www.google.com/patents/WO2016049847A1?cl=en

[34] V. Libal, V. Guralnik, Acquiring statistical access models, US Patent 9,208,156 (Dec. 8 2015). URL https://www.google.com/patents/US9208156

[35] D. S. Xin, J. D. Traupman, X. Meng, P. T. Ogilvie, Dependency management during model compilation of statistical models, US Patent 20,150,379,064 (Dec. 31 2015).

[36] A. Prieston, Method for determining premiums for representation and warranty insurance for mortgage loans, US Patent 8311912 (Nov. 13, 2012). URL https://www.google.com/patents/US8311912

[37] H. Chan, T. Chieu, L. Mok, System to improve predictive maintenance and warranty cost/price estimation, US Patent 8275642 (Sep. 25, 2012). URL https://www.google.com/patents/US8275642

[38] A. Greene, Method of organizing and analysing field warranty data, US Patent 7516175 (Apr. 7, 2009). URL https://www.google.com/patents/US7516175

[39] Y. Chien, N. Choudhury, P. Franklin, 541 S. Kher, H. Rubin, P. Remick, P. Scar , H.Wang, Method and apparatus for warranty cost calculation, US Patent 8131653 (Mar. 6, 2012). URL https://www.google.com/patents/US8131653

[40] T. Rolski, H. Schmidli, V. Schmidt, J. Teugels, Stochastic processes for insurance and finance, Vol. 505, John Wiley & Sons, Chichester, 2009.

[41] K. D. Majeske, A mixture model for automobile warranty data, Reliability Engineering &System Safety 81 (1) (2003) 71-77.

[42] R. A. Ion, V. T. Petkova, B. H. Peeters, P. C. Sander, Field reliability prediction in consumer electronics using warranty data, Quality and Reliability Engineering International 23 (4) (2007) 401-414.

[43] S. Wilson, T. Joyce, E. Lisay, Reliability estimation from field return data, Lifetime data analysis 15 (3) (2009) 397-410.

[44] J. Lawless, J. Hu, J. Cao, Methods for the estimation of failure distributions and rates from automobile warranty data, Lifetime Data Analysis 1 (3) (1995) 227-240.

[45] T. Davis, A simple method for estimating the joint failure time and failure mileage distribution from automobile warranty data, Ford Technical Journal 2 (6) (1999) 1-11.

[46] J. Baik, D. P. Murthy, Reliability assessment based on two-dimensional warranty data and an accelerated failure time model, International Journal of Reliability and Safety 2 (3) (2008) 190-208.

[47] N. D. Singpurwalla, S.Wilson, The warranty problem: its statistical and game-theoretic aspects, SIAM review 35 (1) (1993) 17-42.

[48] H. Moskowitz, Y. H. Chun, A poisson regression model for two-attribute warranty policies, Naval Research Logistics (NRL) 41 (3) (1994) 355-376.

[49] M. Jung, D. Bai, Analysis of field data under two-dimensional warranty, Reliability Engineering & System Safety 92 (2) (2007) 135-143.

[50] S. Wu, Construction of asymmetric copulas and its application in two-dimensional reliability modelling, European Journal of Operational Research 238 (2) (2014) 476-485.

[51] J. Lawless, M. Crowder, K.-A. Lee, Analysis of reliability and warranty claims in products with age and usage scales, Technometrics 51 (1) (2009) 14-24.

[52] J. F. Lawless, M. J. Crowder, Models and estimation for systems with recurrent events and usage processes, Lifetime data analysis 16 (4) (2010) 547-570.

[53] M. Sklar, Fonctions de repartition a n dimensions et leurs marges, Universite Paris 8, Paris, 1959.

[54] P. Shi, E. A. Valdez, Multivariate negative binomial models for insurance claim counts, Insurance: Mathematics and Economics 55 (2014) 18-29.

[55] C. Bolance, Z. Bahraoui, and M. Arts, Quantifying the risk using copulae with nonparametric marginals, Insurance: Mathematics and Economics 58 (2014) 46-56.

[56] D. Diers, M. Eling, S. D. Marek, Dependence modelling in non-life insurance using the Bernstein copula, Insurance: Mathematics and Economics 50 (3) (2012) 430-436.

[57] Y. Zhang, V. Dukic, Predicting multivariate insurance loss payments under the Bayesian copula framework, Journal of Risk and Insurance 80 (4) (2013) 891-919.

[58] M. Fredette, J. Lawless, Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims, Technimetrics 49 (1) (2007) 66-80.

[59] A. Akbarov, S. Wu, Forecasting warranty claims considering dynamic over-dispersion, International Journal of Production Economics 139 (2) (2012) 615-622.

[60] S. Bouguerra, A. Chelbi, N. Rezg, A decision model for adopting an extended warranty under different maintenance policies, International Journal of Production Economics 135 (2) (2012) 840-849.

[61] E. Settanni, L. B. Newnes, N. E. Thenent, G. Parry, Y. M. Goh, A through-life costing methodology for use in product-service-systems, International Journal of Production Economics 153 (2014) 161-177.