Insurance: Mathematics and Economics 87 (2019) 115–129

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

A dependent frequency–severity approach to modeling longitudinal insurance claims

Gee Y. Lee [a], Peng Shi [b,*]

[a] Department of Statistics and Probability, Department of Mathematics, Michigan State University, United States
[b] Wisconsin School of Business, University of Wisconsin-Madison, United States

ARTICLE INFO

Article history:
Received May 2018
Received in revised form April 2019
Accepted 4 April 2019
Available online 18 April 2019

Keywords:
Frequency–severity model
Gaussian copula
Longitudinal insurance claims
Nonlife insurance
Predictive modeling

ABSTRACT

In nonlife insurance, frequency and severity are two essential building blocks in the actuarial modeling of insurance claims. In this paper, we propose a dependent modeling framework to jointly examine the two components in a longitudinal context where the quantity of interest is the predictive distribution. The proposed model accommodates the temporal correlation in both the frequency and the severity, as well as the association between the frequency and severity using a novel copula regression. The resulting predictive claims distribution allows to incorporate the claim history on both the frequency and severity into ratemaking and other prediction applications. In this application, we examine the insurance claim frequencies and severities for specific peril types from a government property insurance portfolio, namely lightning and vehicle claims, which tend to be frequent in terms of their count. We discover that the frequencies and severities of these frequent peril types tend to have a high serial correlation over time. Using dependence modeling in a longitudinal setting, we demonstrate how the prediction of these frequent claims can be improved.

1. Introduction

In nonlife insurance, claims modeling is a critical component in many actuarial applications, including ratemaking, reserving, and claims management. To better quantify the riskiness of policyholders, actuaries often decompose the cost of claims into a frequency component and a severity component. "Frequency" indicates whether a claim has occurred or more generally the number of claims, and "severity" refers to the amount of a claim. In the actuarial literature, this framework is known as the frequency–severity or two-part model. See Klugman et al. (2012) for a discussion in the i.i.d. case and Frees (2014) for a discussion in the regression context. There are several advantages for modeling claim costs using the two-part model. First, the claim cost typically contains a large proportion of zeros corresponding to the policyholders with no claims, and furthermore the positive portion of claim costs are skewed and heavy-tailed. A standard method is not readily available to accommodate these features of the claim data. Second, the frequency and severity components might be affected by different factors. For instance, frequency might be determined mainly by the underlying risk of the policyholder, while severity is to a great extent influenced by the practice of claim adjusters. The two-part framework allows the modeler to incorporate different sets of explanatory variables in a regression setup to account for such differences. Third, separate analyses of frequency and severity provide insurers with useful insights for claims management. Specifically, one could exercise different loss control strategies to mitigate the loss cost with respect to the frequency and severity.

The standard frequency–severity model relies on the independence or conditional independence assumption between the two components, so that the models for the frequency and severity can be implemented separately. This assumption is theoretically appealing because it facilitates the statistical inference for the two processes. However, real insurance claims data have shown that the number and the size of claims of policyholders tend to be correlated. See Frees et al. (2011a) and Erhardt and Czado (2012) for examples in health insurance, and Gschlößl and Czado (2007) and Czado et al. (2012) for examples in automobile insurance.

Motivated by both theoretical developments and empirical evidence, the recent literature has witnessed some development in the strategies to accommodate the dependency between the frequency and the severity in a two-part framework. There are in general two alternative strategies proposed in the literature. The first strand of studies takes a regression approach where the number of claims is treated as an explanatory variable in the regression model for the size of claims. Examples include Gschlößl and Czado (2007), Frees et al. (2011a), Erhardt and Czado (2012), and Garrido et al. (2016) among others. The second strand of studies, aiming to allow for a more flexible dependence structure

* Corresponding author.
E-mail addresses: leegee@msu.edu (G.Y. Lee), pshi@bus.wisc.edu (P. Shi).

https://doi.org/10.1016/j.insmatheco.2019.04.004
0167-6687/

between the frequency and severity components, employs a parametric copula to construct the joint distribution of the number of claims and the average size of claims. For instance, Erhardt and Czado (2012) and Krämer et al. (2013) used a copula approach to jointly model the number and average size of claims for aggregated car insurance data. Shi et al. (2015) adopted a hurdle modeling framework to examine the relationship between the frequency and severity using policy-level claims data. Hua (2015) emphasized the tail dependency between the frequency and severity in the health care utilization context.

In this paper, we further develop this line of literature on the dependent frequency–severity models. We note that the current literature has been limited to cross-sectional insurance claims data, i.e. datasets where both the number and size of claims are collected from a cross-section of policyholders for a given period. For example, Frees et al. (2016) focuses on the cross-sectional dependence among different coverage types of claims. In contrast, our work focuses on the dependent frequency–severity modeling of specific peril types of claims in a predictive modeling context. To be more specific, we assume that one has access to repeated observations on the number and size of claims for policyholders over multiple periods. The quantity of interest is the predictive distribution of the frequency and severity in the future period given past claim history. Because nonlife insurance is in general a short term product, the predictive distribution has wide actuarial applications. For instance, it can provide a mechanism for insurers to incorporate policyholders' claims experience (both frequency and severity) into underwriting and pricing. To this end, we propose a dependent frequency–severity modeling framework for longitudinal insurance claims. The unique feature of longitudinal data is that repeated observations over time tend to be correlated. The proposed model provides a unified framework to capture the serial correlation in the frequency component, the serial correlation in the severity component, as well as the dependence between the frequency and severity. Frequency modeling in a longitudinal context is not unusual in the actuarial literature; see Boucher et al. (2008), Boucher and Guillén (2011), and Shi and Valdez (2014) for examples of recent studies. However, studies that simultaneously consider the frequency and severity in longitudinal settings are rarely found in the literature, not to mention longitudinal models that account for the dependency between the frequency and severity.

In addition to the methodological contribution to the literature, our work provides additional empirical evidence of dependence between the frequency and severity of insurance claims. As indicated above, current evidence that supports a dependent frequency–severity framework is mainly found in automobile and health insurance. In this paper, we examine the insurance claims from the building and contents coverage in property insurance, and find significant correlation between the number of claims and average size of claims in the longitudinal context. More importantly, the association between the frequency and severity along with the association within frequency/severity provide crucial lift in the prediction. To understand the importance of this, note that building and contents are often the largest among the various coverages offered by a nonlife insurance company. The coverage constitutes a major part of both homeowner's property insurance and commercial property insurance. The coverage may also appear in government property insurance programs such as the Local Government Property Insurance Fund (LGPIF) explored in this study. See Cook (2012) and Flitner (2014) for a survey on homeowner's insurance and commercial property insurance, respectively.

The remainder of the paper proceeds in the following order: Section 2 summarizes the LGPIF data, which are used for the empirical analysis. Section 3 introduces the general framework for accommodating the dependence between the frequency and severity in a dynamic setting. Statistical inference including estimation and prediction is presented in Section 4. The data analysis is performed in Section 5, and we demonstrate the superior performance of our dependent model over the independent model in the prediction task in this section. Section 6 concludes the paper with closing remarks.

Table 1
Summary statistics for claim frequency and average severity by peril type.

Year	Lightning frequency				Lightning severity			
	Min	Mean	Max	#obs	Min	Mean	Max	#obs
2006	0	0.157	5	1159	145	11,139	94,523	133
2007	0	0.149	4	1143	513	12,439	224,101	125
2008	0	0.126	4	1130	621	8,987	58,882	114
2009	0	0.118	5	1114	600	11,415	241,049	99
2010	0	0.185	6	1114	623	11,157	270,807	145
2011	0	0.112	4	1096	805	12,664	88,603	96
Year	Vehicle frequency				Vehicle severity			
	Min	Mean	Max	#obs	Min	Mean	Max	#obs
2006	0	0.100	19	1159	425	3,732	21,727	61
2007	0	0.161	18	1143	535	5,251	111,740	89
2008	0	0.141	17	1130	639	3,888	22,433	76
2009	0	0.163	10	1114	287	3,444	24,465	88
2010	0	0.189	13	1114	240	6,619	97,085	96
2011	0	0.207	16	1096	1	6,538	135,268	101

2. Data

We utilize the LGPIF data to explore longitudinal insurance claims modeling for recurrent peril types. The Wisconsin LGPIF had been established to make property insurance available for local government units including cities, counties, schools, villages, towns, and miscellaneous entities including fire stations. The Wisconsin Act 59 allowed for the closure of the LGPIF as of December 31, 2017. Although the fund has closed, the historic data remain a valuable source for academic studies of insurance claims and modeling applications.

In property insurance, claims data are often categorized by perils to indicate the cause of loss. It is often advantageous to model the losses from each peril type separately since different peril types may experience different loss frequencies, severities, and dependencies between the frequencies and severities. The serial correlation over time may also be different for various peril types. Hence, when possible, it helps to model the different peril types separately. Modeling of multi-peril insurance coverages has been explored in Frees et al. (2010), where each claim is treated as a multivariate response wherein the components correspond to the various peril types. In this study, we focus on the lightning and vehicle claims in particular. Lightning and vehicle claims are small frequent claims within the LGPIF building and contents claim category. A lightning strike may cause a fire, ruin electronics, or damage wires inside the walls. If a claim has been initialized by a lightning strike, it is categorized as a lightning claim. Meanwhile, a vehicle claim may be caused by either a car, truck, or a plow running into an insured building and causing damage. Table 1 reports the summary statistics for the frequency and severity of insurance claims from the two perils by year. For frequency, we report the mean, minimum, and maximum number of claims by year. For severity, we report mean, minimum, and maximum amount of claims by year given there is at least one claim during the year. The last column in the table "#obs" shows the number of observations in each year.

Table 2 shows a summary of the explanatory variables used in the model. The explanatory variables are common for both

Table 2
Description and sample mean of rating variables.

Variable	Description	Mean						
		2006	2007	2008	2009	2010	2011	Overall
City	Indicator of city	0.140	0.139	0.140	0.143	0.140	0.141	0.140
County	Indicator of county	0.053	0.054	0.055	0.064	0.064	0.065	0.059
Misc	Indicator of misc. entities	0.108	0.106	0.109	0.108	0.110	0.115	0.109
School	Indicator of school	0.282	0.285	0.283	0.282	0.280	0.276	0.282
Town	Indicator of town	0.184	0.175	0.171	0.170	0.167	0.165	0.172
Village	Indicator of village	0.233	0.241	0.242	0.234	0.239	0.238	0.238
AC05	Indicator of 5% alarm credit	0.024	0.025	0.036	0.053	0.074	0.074	0.047
AC10	Indicator of 10% alarm credit	0.044	0.050	0.050	0.065	0.081	0.084	0.062
AC15	Indicator of 15% alarm credit	0.367	0.388	0.421	0.475	0.533	0.538	0.452
HighFreq	Indicator of high frequency	0.003	0.003	0.004	0.004	0.004	0.004	0.004
lnDeductBC	Log deductible amount	7.064	7.126	7.158	7.209	7.219	7.241	7.169
lnCoverageBC	Log coverage amount	1.951	2.064	2.134	2.224	2.232	2.267	2.143

Table 3
Number of policyholders over time, by entity type and by alarm credit.

Year	Entity type						Alarm credit				Total
	City	County	Misc.	School	Town	Village	None	AC05	AC10	AC15	
2006	162	62	125	327	213	270	655	28	51	425	1159
2007	159	62	121	326	200	275	614	29	57	443	1143
2008	158	62	123	320	193	274	556	41	57	476	1130
2009	159	71	120	314	189	261	454	59	72	529	1114
2010	156	71	123	312	186	266	348	82	90	594	1114
2011	154	71	126	303	181	261	333	81	92	590	1096

Table 4
Descriptive statistics of frequency and severity by deductible and by peril type.

Deduct	Pr. Non-zero loss			Severity			
	Lightning	Vehicle	#obs	Lightning	#losses	Vehicle	#losses
500	0.103	0.074	3132	7,413	324	3,422	233
1 000	0.119	0.091	1314	9,718	156	3,821	119
2 500	0.107	0.057	826	20,870	88	6,654	47
5 000	0.085	0.059	867	16,839	74	10,564	51
10 000	0.045	0.053	247	12,798	11	14,370	13
15 000	0.086	0.049	81	52,216	7	7,833	4
25 000	0.170	0.143	224	7,766	38	4,825	32
50 000	0.211	0.237	38	20,861	8	10,007	9
75 000	1.000	0.500	6	9,297	6	7,707	3

perils, and the table presents the overall mean during 2006–2011 as well as the sample mean by year. City, County, Misc, School, Town, and Village are indicator variables of the entity type. Note that entities in this dataset are local government policyholders with a number of buildings to insure under the coverage. AC05 is an indicator of whether at least one of the buildings receives a 5% alarm credit. A 5% alarm credit means the entity receives a 5% discount in premium if automatic smoke alarms are installed in some of the main rooms within a building. The entity receives a 10% discount if alarms are installed in all of the main rooms. Finally, the entity receives a 15% discount if the alarms are monitored 24 h, 7 days a week. Some of the entities have extraordinarily high frequencies, and in order to treat these policies separately, we use a binary variable, HighFreq. Those policies with HighFreq equal to 1 have had at least one year in which the number of building and contents claims was greater than 50. The reason why this was done is to treat outliers separately. Note that the number of entities with HighFreq equal to 1 is very small; only 0.4% of the entities are in this category. The rest of the variables are self-explanatory. lnDeductBC is the log deductible amount, and lnCoverageBC is the log coverage amount in millions of dollars. The coverage amount can be considered as the maximum possible claim amount.

From Table 1 we can see that around one thousand entities incur more than 0.1 claims each year. The number of entities changes over time, as indicated by Table 3. Although the total number of entities has decreased over time, as indicated by the right most column of the table, some entities have increased in number over time. For example, the county entity had 62 entities until year 2008, and then there are 71 counties starting in year 2009. The number of entities with alarm credit is shown in the right panel of Table 3. The number of entities with 15% alarm credit seems to be increasing over time, while those with no alarm credit decrease over time. Presumably, this is because more alarms are installed within the buildings as time passes by.

Table 4 provides some deductible level summary statistics to give an intuitive understanding of the effect of the deductible rating variable. Specifically, we report the proportion of observations with no losses, and the average loss amount for the losses that do occur, by each deductible level. We observe that policyholders tend to select smaller deductibles, and hence the largest number of observations occur in the 500 deductible category. In general, the proportion of non-zero losses tends to be large for those policyholders who selected large deductible levels. Note that the smaller sample size suggests higher uncertainty in the proportion of non-zero losses and the average severities associated with larger deductibles.

The average severity amounts shown in Table 4 are the underlying loss amounts, as opposed to the censored claims. The relationship between the deductible and the response variable may be modeled using two different approaches, namely the regression approach and the maximum likelihood approach. In this paper, we are interested in modeling the underlying losses while allowing for serial dependence. Thus, we will treat the deductible as an explanatory variable in a regression model and, more specifically, use the log-deductible amount as a rating variable. See Lee (2017) for more discussion on the treatment of deductibles in claims modeling.

Tables 5 and 6 present the serial correlation for lightning and vehicle claims for the frequency and severity, respectively. Due to the discreteness of frequency, we use the polychoric correlation described in Joe (2014). Both lightning and vehicle perils show strong serial correlation in the number of claims. Spearman's correlation is reported for the average claim severity. The relationship for the claim severity is slightly weaker compared to that for frequency. Nonetheless, there seems to exist a moderate correlation over time.

We also computed the polyserial correlation (see Joe (2014)) between the frequencies and severities of the lightning claims and vehicle claims. We discovered that the polyserial correlation for the lightning claims is 0.004 whereas the polyserial correlation for the vehicle claims is -0.104. This indicates that the dependence between the frequencies and severities may not be strong for the lightning claims, but strong for the vehicle claims.

Table 5
Polychoric correlation for frequencies of lightning and vehicle claims.

	Lightning					Vehicle				
	2006	2007	2008	2009	2010		2006	2007	2008	2009	2010
2007	0.501					2007	0.766				
2008	0.518	0.518				2008	0.774	0.822			
2009	0.517	0.546	0.377			2009	0.705	0.766	0.768		
2010	0.555	0.532	0.622	0.557		2010	0.707	0.767	0.721	0.782	
2011	0.379	0.502	0.524	0.454	0.589	2011	0.668	0.778	0.728	0.774	0.775

Table 6
Spearman's correlation for severities of lightning and vehicle claims.

	Lightning					Vehicle					
	2006	2007	2008	2009	2010		2006	2007	2008	2009	2010
2007	0.241					2007	0.343				
2008	0.246	0.264				2008	0.344	0.420			
2009	0.248	0.248	0.165			2009	0.305	0.393	0.355		
2010	0.285	0.243	0.354	0.265		2010	0.326	0.387	0.339	0.400	
2011	0.164	0.260	0.256	0.232	0.307	2011	0.263	0.404	0.339	0.348	0.380

3. Methodology

3.1. A general framework

We examine the collective risk model in a longitudinal setup. Let S_{it} denote the aggregate claim cost for policyholder $i \in \{1,\ldots,m\}$ in period $t \in \{1,\ldots,T\}$. The collective risk model defines $S_{it} = Z_{it,1} + \cdots + Z_{it,N_{it}}$, where $Z_{it,n_{it}}$ denotes the size of the n_{it}th claim for $n_{it} = 1,\ldots,N_{it}$. We reformulate the model as follows:

$$S_{it} = N_{it} \times Y_{it}, \text{ where } Y_{it} = \begin{cases} S_{it}/N_{it}, & N_{it} > 0 \\ 0, & N_{it} = 0 \end{cases}. \tag{1}$$

We call N_{it} frequency and Y_{it} (average) severity. It is straightforward to see that N_{it} and Y_{it} are not independent even when N_{it} and Z_{it} are independent because Y_{it} is a function of N_{it}. In fact, one can show that N_{it} and Y_{it} are positively correlated due to the mass at zero. In this work, we instead focus on the relationship between N_{it} and Y_{it} given $N_{it} > 0$, and we allow the data to depict this relationship. Interestingly, the frequency and severity turn out to be negatively dependent, as will be shown by the empirical results in Section 5.

For compact notations, define $\boldsymbol{N}_i = (N_{i1},\ldots,N_{iT})$ and $\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{iT})$, and let $\boldsymbol{n}_i$ and $\boldsymbol{y}_i$ denote their realizations, respectively. We propose the joint model

$$f_{\boldsymbol{N},\boldsymbol{Y}}(\boldsymbol{n}_i,\boldsymbol{y}_i) = f_{\boldsymbol{N}}(\boldsymbol{n}_i) \times f_{\boldsymbol{Y}|\boldsymbol{N}}(\boldsymbol{y}_i|\boldsymbol{n}_i) \tag{2}$$

where $f_{\boldsymbol{N},\boldsymbol{Y}}$ denotes the joint distribution of $(\boldsymbol{N},\boldsymbol{Y})$, $f_{\boldsymbol{N}}$ denotes the joint pmf of $\boldsymbol{N}$, and $f_{\boldsymbol{Y}|\boldsymbol{N}}$ denotes the joint pmf/pdf of $\boldsymbol{Y}$ conditional on $\boldsymbol{N}$. We emphasize two observations in this formulation. First, the distribution of $\boldsymbol{Y}$ given $\boldsymbol{N}$ is mixed because Y_t is continuous for $N_t > 0$ and degenerate for $N_t = 0$. Second, Eq. (2) is a result of conditional probability and thus does not require any additional constraint to be valid.

We use parametric copulas to construct each of the two components in Eq. (2). A copula is a general model for constructing multivariate distributions, and it has found extensive applications in the actuarial literature. See Nelsen (1999) for a mild introduction and Joe (2014) for a comprehensive review on the most recent developments. For the frequency, the joint distribution can be represented using a T-variate copula C_T^{Freq} as

$$\begin{aligned} F_{\boldsymbol{N}}(\boldsymbol{n}_i) &= \Pr(N_{i1} \le n_{i1},\ldots,N_{iT} \le n_{iT}) \\ &= C_T^{\mathrm{Freq}}(F_{N_1}(n_{i1}),\ldots,F_{N_T}(n_{iT})). \end{aligned} \tag{3}$$

To derive the associated probability mass function, one has

$$\begin{aligned} f_{\boldsymbol{N}}(\boldsymbol{n}_i) &= \Pr(N_{i1} = n_{i1},\ldots,N_{iT} = n_{iT}) \\ &= \sum_{l_1=0}^{1}\cdots\sum_{l_T=0}^{1}(-1)^{l_1+\cdots+l_T} \times C_T^{\mathrm{Freq}}(u_{i1,l_1},\ldots,u_{iT,l_T}) \end{aligned} \tag{4}$$

where $u_{it,0} = F_{N_t}(n_{it})$ and $u_{it,1} = F_{N_t}(n_{it}-1)$.

The joint distribution of the conditional severity given frequency is formulated using another T-variate copula C_T^{Sev} as

$$\begin{aligned} F_{\boldsymbol{Y}|\boldsymbol{N}}(\boldsymbol{y}_i|\boldsymbol{n}_i) &= \Pr(Y_{i1} \le y_{i1},\ldots,Y_{iT} \le y_{iT}|\boldsymbol{N}_i = \boldsymbol{n}_i) \\ &= C_T^{\mathrm{Sev}}(F_{Y_1|\boldsymbol{N}}(y_{i1}|\boldsymbol{n}_i),\ldots,F_{Y_T|\boldsymbol{N}}(y_{iT}|\boldsymbol{n}_i)) \\ &= C_T^{\mathrm{Sev}}(F_{Y_1|N_1}(y_{i1}|n_{i1}),\ldots,F_{Y_T|N_T}(y_{iT}|n_{iT})) \end{aligned} \tag{5}$$

where we assume that the copula C_T^{Sev} does not depend on the value of the frequency. In addition, given the frequency of the current period, the severity does not depend on the previous frequency, i.e. $F_{Y_t|\boldsymbol{N}}(y_{it}|\boldsymbol{n}_i) = F_{Y_t|N_t}(y_{it}|n_{it})$. As pointed out earlier, the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{N}$ is mixed, because some components of $\boldsymbol{Y}$ are discrete with probability mass of one on zero, and some components of $\boldsymbol{Y}$ are continuous, depending on the associated claim frequency. Let $t_i^+ = \{t : n_{it} \neq 0, t = 1,\ldots,T\} = \{t_{ij} : j = 1,\ldots,J_i\}$ denote the set of times when non-zero severities are observed. Then we express the joint pmf/pdf of $\boldsymbol{Y}$ given $\boldsymbol{N}$ as

$$\begin{aligned} &f_{\boldsymbol{Y}|\boldsymbol{N}}(\boldsymbol{y}_i|\boldsymbol{n}_i) \\ &= \frac{\partial^{J_i}}{\partial y_{it_{i1}}\cdots\partial y_{it_{iJ_i}}}\Pr(Y_{it_{i1}} \le y_{it_{i1}},\ldots,Y_{it_{iJ_i}} \le y_{it_{iJ_i}}, \boldsymbol{Y}_{i\overline{t_i^+}} = \boldsymbol{0}|N_{it_{i1}} \\ &\qquad = n_{it_{i1}},\ldots,N_{it_{iJ_i}} = n_{it_{iJ_i}}, \boldsymbol{N}_{i\overline{t_i^+}} = \boldsymbol{0}) \\ &= \prod_{j=1}^{J_i} f_{Y_t|N_t}(y_{it_{ij}}|n_{it_{ij}}) \cdot c^{\mathrm{Sev}}_{T:t_{i1},\ldots,t_{iJ_i}}(\omega_{i1},\ldots,\omega_{iT}) \end{aligned} \tag{6}$$

where $\overline{t_i^+}$ denotes the complement set of t_i^+,

$$\omega_{it} = \begin{cases} F_{Y_t|N_t}(y_{it}|n_{it}) & t \in t_i^+ \\ 1 & \text{otherwise,} \end{cases} \tag{7}$$

and $c^{\mathrm{Sev}}_{T:t_{i1},\ldots,t_{iJ_i}}$ is defined as

$$c^{\mathrm{Sev}}_{T:t_{i1},\ldots,t_{iJ_i}}(u_1,\ldots,u_T) = \frac{\partial^{J_i}}{\partial u_{t_{i1}}\cdots\partial u_{t_{iJ_i}}} C_T^{\mathrm{Sev}}(u_1,\ldots,u_T). \tag{8}$$

To finalize (6), one needs the conditional distribution of severity given frequency. This conditional distribution is derived from the joint distribution of (N_t, Y_t). To do so, we propose using a bivariate copula C^{FS} to join the frequency and severity components, so that their joint distribution becomes

$$\begin{aligned} F_{N_t,Y_t}(n_{it},y_{it}) &= \Pr(N_{it} \le n_{it}, Y_{it} \le y_{it}) \\ &= \begin{cases} F_{N_t}(0) & \text{if } n_{it} = 0 \\ (1-F_{N_t}(0)) \cdot F_{N_t,Y_t}(n_{it},y_{it}|N_{it} > 0) & \text{if } n_{it} > 0 \end{cases} \\ &= \begin{cases} F_{N_t}(0) & \text{if } n_{it} = 0 \\ (1-F_{N_t}(0)) \cdot C^{\mathrm{FS}}\left(F_{N_t}(n_{it}|N_{it} > 0), F_{Y_t}(y_{it}|N_{it} > 0)\right) & \text{if } n_{it} > 0 \end{cases}. \end{aligned} \tag{9}$$

Note that in the above, C^{FS} is not the copula for the joint distribution of (N_{it}, Y_{it}), rather it corresponds to the copula that joins the frequency and severity conditional on the number of claims being greater than zero. Hence, the conditional distribution of the severity given the frequency can be written as Eq. (10) in Box I, where $\boldsymbol{1}_A$ is an indicator function for event A, $u^+_{it,0} = F_{N_t}(n_{it}|N_{it} > 0)$ and $u^+_{it,1} = F_{N_t}(n_{it}-1|N_{it} > 0)$. And the corresponding probability density or mass function of the conditional severity

$$F_{Y_t|N_t}(y_{it}|n_{it}) = \frac{F_{N_t,Y_t}(n_{it},y_{it}) - F_{N_t,Y_t}(n_{it}-1,y_{it})}{f_{N_t}(n_{it})} \tag{10}$$
$$= \begin{cases} \mathbf{1}_{\{y_{it}\le 0\}} & \text{if } n_{it}=0 \\ \dfrac{(1-F_{N_t}(0))\cdot \sum_{l_t=0}^{1}(-1)^{l_t} C^{\mathrm{FS}}\left(u^{+}_{it,l_t}, F_{Y_t}(y_{it}|N_{it}>0)\right)}{f_{N_t}(n_{it})} & \text{if } n_{it}>0 \end{cases},$$

Box I.

$$f_{Y_t|N_t}(y_{it}|n_{it}) \tag{11}$$
$$= \begin{cases} \mathbf{1}_{\{y_{it}=0\}} & \text{if } n_{it}=0 \\ \dfrac{(1-F_{N_t}(0))\cdot f_{Y_t}(y_{it}|N_{it}>0)\cdot \sum_{l_t=0}^{1}(-1)^{l_t} c_2^{\mathrm{FS}}\left(u^{+}_{it,l_t}, F_{Y_t}(y_{it}|N_{it}>0)\right)}{f_{N_t}(n_{it})} & \text{if } n_{it}>0 \end{cases},$$

Box II.

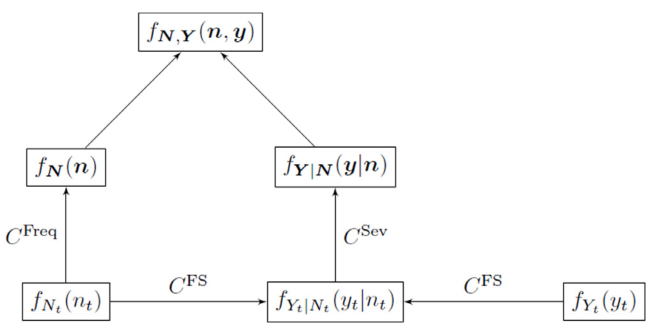

Fig. 1. Flow chart for the model building process.

is Eq. (11) in Box II, where

$$c_2^{\mathrm{FS}}(u_1,u_2) = \frac{\partial}{\partial u_2} C^{\mathrm{FS}}(u_1,u_2).$$

Combining (4) and (6), one has the joint distribution of $(\boldsymbol{N},\boldsymbol{Y})$ as defined by (2), where one requires (10) and (11) in the evaluation of (6). Note that in the entire model building process, we require the simplifying assumption that the three copulas C_T^{Freq}, C_T^{Sev}, and C^{FS} have constant association parameters. The framework does not require any additional assumption on the dependence structure, and thus, at least theoretically, the copulas are not limited to a specific form such as elliptical or Archimedean.

Example

To clarify the notations, let us consider an example of $T=3$. The joint distribution of $(\boldsymbol{N}_i,\boldsymbol{Y}_i) = (N_{i1},N_{i2},N_{i3},Y_{i1},Y_{i2},Y_{i3})$ is expressed as

$$f_{\boldsymbol{N},\boldsymbol{Y}}(n_{i1},n_{i2},n_{i3},y_{i1},y_{i2},y_{i3}) = f_{\boldsymbol{N}}(n_{i1},n_{i2},n_{i3}) \times f_{\boldsymbol{Y}|\boldsymbol{N}}(y_{i1},y_{i2},y_{i3}|n_{i1},n_{i2},n_{i3}),$$

where

$$\begin{aligned} &f_{\boldsymbol{N}}(n_{i1},n_{i2},n_{i3}) \\ =&C_3^{\mathrm{Freq}}(F_1(n_{i1}),F_2(n_{i2}),F_3(n_{i3})) - C_3^{\mathrm{Freq}}(F_1(n_{i1}-1),F_2(n_{i2}),F_3(n_{i3})) \\ &- C_3^{\mathrm{Freq}}(F_1(n_{i1}),F_2(n_{i2}-1),F_3(n_{i3})) \\ &- C_3^{\mathrm{Freq}}(F_1(n_{i1}),F_2(n_{i2}),F_3(n_{i3}-1)) \\ &+ C_3^{\mathrm{Freq}}(F_1(n_{i1}-1),F_2(n_{i2}-1),F_3(n_{i3})) \\ &+ C_3^{\mathrm{Freq}}(F_1(n_{i1}-1),F_2(n_{i2}),F_3(n_{i3}-1)) \\ &+ C_3^{\mathrm{Freq}}(F_1(n_{i1}),F_2(n_{i2}-1),F_3(n_{i3}-1)) \\ &- C_3^{\mathrm{Freq}}(F_1(n_{i1}-1),F_2(n_{i2}-1),F_3(n_{i3}-1)) \end{aligned}$$

and

$$f_{\boldsymbol{Y}|\boldsymbol{N}}(y_{i1},y_{i2},y_{i3}|n_{i1},n_{i2},n_{i3})$$
$$= \begin{cases} 1 & \text{if } t_i^+ = \emptyset \\ f_{Y_1|N_1}(y_{i1}|n_{i1}) c^{Sev}_{3:1}\left(F_{Y_1|N_1}(y_{i1}|n_{i1}),1,1\right) & \text{if } t_i^+ = \{1\} \\ f_{Y_2|N_2}(y_{i2}|n_{i2}) c^{Sev}_{3:2}\left(1,F_{Y_2|N_2}(y_{i2}|n_{i2}),1\right) & \text{if } t_i^+ = \{2\} \\ f_{Y_3|N_3}(y_{i3}|n_{i3}) c^{Sev}_{3:3}\left(1,1,F_{Y_3|N_3}(y_{i3}|n_{i3})\right) & \text{if } t_i^+ = \{3\} \\ f_{Y_1|N_1}(y_{i1}|n_{i1}) f_{Y_2|N_2}(y_{i2}|n_{i2}) c^{Sev}_{3:1,2} & \\ (F_{Y_1|N_1}(y_{i1}|n_{i1}),F_{Y_2|N_2}(y_{i2}|n_{i2}),1) & \text{if } t_i^+ = \{1,2\} \\ f_{Y_1|N_1}(y_{i1}|n_{i1}) f_{Y_3|N_3}(y_{i3}|n_{i3}) c^{Sev}_{3:1,3} & \\ (F_{Y_1|N_1}(y_{i1}|n_{i1}),1,F_{Y_3|N_3}(y_{i3}|n_{i3})) & \text{if } t_i^+ = \{1,3\} \\ f_{Y_2|N_2}(y_{i2}|n_{i2}) f_{Y_3|N_3}(y_{i3}|n_{i3}) c^{Sev}_{3:2,3} & \\ (1,F_{Y_2|N_2}(y_{i2}|n_{i2}),F_{Y_3|N_3}(y_{i3}|n_{i3})) & \text{if } t_i^+ = \{2,3\} \\ \prod_{t=1}^{3} f_{Y_t|N_t}(y_{it}|n_{it}) c^{Sev}_{3:1,2,3}(F_{Y_1|N_1}(y_{i1}|n_{i1}), & \\ F_{Y_2|N_2}(y_{i2}|n_{i2}),F_{Y_3|N_3}(y_{i3}|n_{i3})) & \text{if } t_i^+ = \{1,2,3\} \end{cases},$$

where

$$F_{Y_t|N_t}(y_{it}|n_{it}) = (1-F_{N_t}(0)) \times \frac{C^{\mathrm{FS}}\left(u^+_{it,0},F_{Y_t}(y_{it}|N_{it}>0)\right) - C^{\mathrm{FS}}\left(u^+_{it,1},F_{Y_t}(y_{it}|N_{it}>0)\right)}{f_{N_t}(n_{it})},$$

and $u^+_{it,0} = F_{N_t}(n_{it}|N_{it}>0)$ and $u^+_{it,1} = F_{N_t}(n_{it}-1|N_{it}>0)$, for $t=1,2,3$.

Summary

As a summary, the proposed model uses three copulas to accommodate the dependencies in the complex claims data. The copula C_T^{Freq} is employed to account for the temporal association

Table 7
Summary of notations and abbreviations.

Symbol	Description
S_{it}	The aggregate loss cost
N_{it}	The number of claims
Y_{it}	The average claim amount
f_{N_t}, F_{N_t}	The pmf and cdf of claim frequency
f_{Y_t}, F_{Y_t}	The pdf and cdf of average claim severity
F_{N_t,Y_t}	The joint distribution of frequency and severity
$F_{Y_t\|N_t}$	The conditional distribution of severity given frequency
$u_{it,0}$	$F_{N_t}(n_{it})$
$u_{it,1}$	$F_{N_t}(n_{it}-1)$
$u^{+}_{it,0}$	$F_{N_t}(n_{it}\|N_{it}>0)$
$u^{+}_{it,1}$	$F_{N_t}(n_{it}-1\|N_{it}>0)$
C_T^{Freq}	The copula for the temporal dependence in frequency
C^{Sev}	The copula for the temporal dependence in conditional severity given frequency
C_T^{FS}	The copula for the cross-sectional relation between frequency and severity
ω_{it}	$F_{Y_t\|N_t}(y_{it}\|n_{it})$ for $n_{it}>0$ and 1 for $n_{it}=0$

in claim frequency, the copula C^{FS} is used to capture the cross-sectional relation between the frequency and severity, and the copula C_T^{Sev} is used to accommodate the temporal association in average severities given frequency. A flowchart is provided in Fig. 1 to visualize the steps in model formulation. In the figure, each box represents a building block, and each directed edge indicates the flow of the model building process. The copula on the edge connects the two building blocks in the nodes.

The model building process of the proposed copula model is based on a set of key assumptions which are summarized in the list below:

(i) Given the frequency of current period N_{it}, the severity Y_{it} does not depend on the previous frequency N_{is} for $s<t$;
(ii) The association parameters in copula C_T^{Freq} do not depend on explanatory variables;
(iii) The association parameters in copula C_T^{Sev} do not depend on explanatory variables nor the claim frequency;
(iv) The copula C^{FS} is stationary over time.

Notations and abbreviations used throughout the paper are summarized in Table 7 to help the reader to follow the paper.

3.2. Model specification

3.2.1. Marginal model

For claim frequency, due to the excessive number of zeros, analysts often use zero-inflated count regression models as in Boucher (2014). In this application, we consider the class of zero–one-inflated models as in Frees et al. (2016). Similar to the zero-inflated model, the zero–one-inflation model employs two generating processes. The zero–one-inflated model extends the zero-inflated method in that a separate generating process is used for both the zeros and ones. To be more specific, the first process is governed by a multinomial distribution that generates structural zeros and ones. The second process is governed by a standard count regression model.

Denote the latent variable in the first process as I_{it}, which follows a multinomial distribution with possible values 0, 1 and 2 and the associated probabilities $\pi_{0,it}, \pi_{1,it}, \pi_{2,it} = 1-\pi_{0,it}-\pi_{1,it}$. Let $P_{it}(n)$ be the probability mass function for the standard count distribution in the second process. Then the probability mass function of N_{it} can be expressed as

$$f_{N_t}(n) = \pi_{0,it}\mathbf{1}_{\{n=0\}} + \pi_{1,it}\mathbf{1}_{\{n=1\}} + \pi_{2,it}P_{it}(n), \tag{12}$$

Let $\boldsymbol{x}_{it}$ denote the covariates that one could use to account for observed heterogeneity. We employ a logit specification to parameterize the probabilities for the latent variable I_{it}. Using level 2 as the reference category, the specification is

$$\ln\frac{\pi_{j,it}}{\pi_{2,it}} = \boldsymbol{x}_{it}'\boldsymbol{\gamma}_j, j=0,1.$$

Correspondingly,

$$\pi_{j,it} = \frac{\exp(\boldsymbol{x}_{it}'\boldsymbol{\gamma}_j)}{1+\exp(\boldsymbol{x}_{it}'\boldsymbol{\gamma}_0)+\exp(\boldsymbol{x}_{it}'\boldsymbol{\gamma}_1)}, \quad j=0,1$$
$$\pi_{2,it} = 1-\pi_{0,it}-\pi_{1,it}.$$

The distribution $P_{it}(n)$ is specified using a negative binomial model with a log link.

For severity, we specify the distribution of average claim amount given there is at least one claim. We employ the generalized beta of the second kind, or in short GB2, distribution to account for the skewness and heavy tails in the claim amounts. The density of the GB2 distribution is

$$f_{Y_t}(y|N_{it}>0) = \frac{|a|(y/b)^{a\cdot\alpha_1}}{yB(\alpha_1,\alpha_2)[1+(y/b)^a]^{\alpha_1+\alpha_2}}, \tag{13}$$

where $B(\alpha_1,\alpha_2)$ is the beta function. The GB2 distribution is defined for $0<y_{it}<\infty$, and it belongs to the family of generalized beta distribution. See Shi (2014) for an overview of various alternative approaches to claim severity modeling.

For the purpose of incorporating covariates in a regression context, we consider the following reparameterizations:

$$f_{Y_t}(y|N_{it}>0) = \frac{[\exp(z)]^{\alpha_1}}{y_{it}\sigma B(\alpha_1,\alpha_2)[1+\exp(z)]^{\alpha_1+\alpha_2}}, \tag{14}$$

where $z = (\ln(y)-\mu_{it})/\sigma$, and the location parameter is further specified as a linear combination of covariates, i.e. $\mu_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}$.

3.2.2. Dependence

For the dependence modeling, we consider the normal copula. Let $\boldsymbol{\Sigma}$ denote the correlation matrix. Then the multivariate normal copula is

$$C(\boldsymbol{u};\boldsymbol{\Sigma}) = \Phi_p(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_p);\boldsymbol{\Sigma}), \boldsymbol{u}\in[0,1]^p \tag{15}$$

and the corresponding copula density is

$$c(\boldsymbol{u};\boldsymbol{\Sigma}) = \frac{\phi_p(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_p);\boldsymbol{\Sigma})}{\prod_{j=1}^p\phi(\Phi^{-1}(u_j))} \tag{16}$$

where Φ is the distribution function for the standard normal random variable, Φ_p is the distribution for a p-dimensional normal vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, and ϕ and ϕ_p are the respective associated density functions.

In this work, we need three copulas for accommodating the temporal dependence in frequency, the temporal dependence in severity, and the contemporaneous dependence between the frequency and severity. All three copulas are specified as normal copulas. We consider the following correlation matrix for copulas C_T^{Freq}, C^{FS}, and C_T^{Sev}, respectively:

$$\boldsymbol{\Sigma}^{\text{Freq}} = \begin{bmatrix} 1 & \rho_{Freq} & \ldots & \rho_{Freq}^{T-1} \\ \rho_{Freq} & 1 & \ldots & \rho_{Freq}^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{Freq}^{T-1} & \rho_{Freq}^{T-2} & \ldots & 1 \end{bmatrix}, \boldsymbol{\Sigma}^{\text{FS}} = \begin{bmatrix} 1 & \rho_{FS} \\ \rho_{FS} & 1 \end{bmatrix},$$

$$\boldsymbol{\Sigma}^{\mathrm{Sev}}=\begin{bmatrix}1 & \rho_{Sev} & \dots & \rho_{Sev}^{T-1}\\ \rho_{Sev} & 1 & \dots & \rho_{Sev}^{T-2}\\ \vdots & \vdots & \ddots & \vdots\\ \rho_{Sev}^{T-1} & \rho_{Sev}^{T-2} & \dots & 1\end{bmatrix}.$$

There are several advantages of using the Gaussian copula in the proposed model. First, it is simple and straightforward to use in applications, and the properties of Gaussian copulas have been well studied in the literature. Second, the correlation matrix allows for flexible yet interpretable dependence for structured data. For instance, the stationary serial correlation commonly used in time series data such as AR1 can be easily implemented in the Gaussian copula framework. Third, the Gaussian copula is readily able to accommodate unbalanced data in longitudinal studies because it is complete under marginalization. In this study, the lack of balance occurs in two cases: (1) The frequency data could be unbalanced if a policyholder is not observed for all sampling periods; (2) The severity data are usually unbalanced unless a policyholder has claims in all observation periods, a rare situation in property insurance. Lastly, the Gaussian copula is particularly valuable for predictions, where one requires a dependence structure between the future outcome and the past observations. We note that any other member of the elliptical copula family could be used in the proposed framework. However, our experience suggests that the extra complexity usually does not add much value for the kind of applications found in this paper.

Although the Gaussian copula is well motivated for this particular work, we are aware of its potential limitations in other applications. First, the computational burden of working with the Gaussian copula for discrete data could be large for the high dimensional case. An alternative is the pair copula construction approach proposed by Panagiotelis et al. (2012), although its application to unbalanced data is not as straightforward as the Gaussian copula. Second, the Gaussian copula cannot capture tail dependence. When modeling granular level claim data in our work, we did not find tail dependence critical for the prediction. In addition, we emphasize that there are certainly different strategies to specify the multivariate copula in different parts of the model; the main contribution of our work is to introduce a general framework to study frequency–severity dependence in a longitudinal setting, and the copula construction is secondary.

4. Statistical inference

4.1. Estimation

The parameters in the proposed longitudinal data model can be estimated using the likelihood-based approach. Define the parameter vector $\boldsymbol{\Lambda} = \left(\boldsymbol{\theta}^{\mathrm{Freq}'}, \boldsymbol{\theta}^{\mathrm{Sev}'}, \boldsymbol{\zeta}^{\mathrm{Freq}'}, \boldsymbol{\zeta}^{\mathrm{Sev}'}, \boldsymbol{\zeta}^{\mathrm{FS}'}\right)'$, where $\boldsymbol{\theta}$ denotes the parameters in the marginal distributions, and $\boldsymbol{\zeta}$ denote the parameters in the copulas. The log likelihood function is

$$l(\boldsymbol{\Lambda})=\sum_{i=1}^{m} l_i(\boldsymbol{\Lambda})=\sum_{i=1}^{m}\left\{\ln f_{\boldsymbol{N}}(\boldsymbol{n}_i)+\ln f_{\boldsymbol{Y}|\boldsymbol{N}}(\boldsymbol{y}_i|\boldsymbol{n}_i)\right\}. \tag{17}$$

To gain computational efficiency, we employ the inference functions for margins (IFM) in the likelihood-based estimation. The IFM is a special case of stage-wise estimation technique and has been widely used for estimating copula-based dependence models. Refer to Joe (2014) for a detailed discussion on the IFM method for copula models. To be more specific, the IFM involves two steps. The first step estimates the parameters in the marginal regression models assuming all copulas are independence copulas. Let $\boldsymbol{\Lambda}_1 = \left(\boldsymbol{\theta}^{\mathrm{Freq}'}, \boldsymbol{\theta}^{\mathrm{Sev}'}\right)'$. For the proposed frequency–severity model, the first step estimates $\boldsymbol{\Lambda}_1$ by maximizing:

$$l(\boldsymbol{\Lambda}_1)=\sum_{i=1}^{m} l_i(\boldsymbol{\Lambda}_1)=\sum_{i=1}^{m}\left[\ell_i(\boldsymbol{\theta}^{\mathrm{Freq}})+\ell_i(\boldsymbol{\theta}^{\mathrm{Sev}})\right], \tag{18}$$

where

$$\ell_i(\boldsymbol{\theta}^{\mathrm{Freq}})=\sum_{t=1}^{T}\ln f_{N_t}(n_{it}),$$
$$\ell_i(\boldsymbol{\theta}^{\mathrm{Sev}})=\sum_{\{t:n_{it}>0\}}\ln f_{Y_t}(y_{it}|N_{it}>0).$$

Note that $\hat{\boldsymbol{\Lambda}}_1$ which maximizes (18) can be found by maximizing $\sum_{i=1}^{m}\ell_i(\boldsymbol{\theta}^{\mathrm{Freq}})$ and $\sum_{i=1}^{m}\ell_i(\boldsymbol{\theta}^{\mathrm{Sev}})$ separately. The second step estimates the parameters in the copulas while holding the estimates $\hat{\boldsymbol{\Lambda}}_1$ from the first step fixed. Let $\boldsymbol{\Lambda}_2 = \left(\boldsymbol{\zeta}^{\mathrm{Freq}'}, \boldsymbol{\zeta}^{\mathrm{Sev}'}, \boldsymbol{\zeta}^{\mathrm{FS}'}\right)'$. For the proposed copula model, the second stage estimates $\boldsymbol{\Lambda}_2$ by maximizing

$$l(\boldsymbol{\Lambda}_2)=\sum_{i=1}^{m} l_i(\boldsymbol{\Lambda}_2)=\sum_{i=1}^{m}\left[\ell_i(\boldsymbol{\zeta}^{\mathrm{Freq}})+\ell_i(\boldsymbol{\zeta}^{\mathrm{Sev}},\boldsymbol{\zeta}^{\mathrm{FS}})\right], \tag{19}$$

where

$$\begin{aligned}\ell_i(\boldsymbol{\zeta}^{\mathrm{Freq}})&=\ln f_{\boldsymbol{N}}(\boldsymbol{n}_i)\\&=\ln\left\{\sum_{l_1=0}^{1}\cdots\sum_{l_T=0}^{1}(-1)^{l_1+\cdots+l_T}\right.\\&\qquad\left.\times C_T^{\mathrm{Freq}}(\hat{u}_{i1,l_1},\dots,\hat{u}_{iT,l_T})\right\},\end{aligned}$$

$$\begin{aligned}\ell_i(\boldsymbol{\zeta}^{\mathrm{Sev}},\boldsymbol{\zeta}^{\mathrm{FS}})&=\ln f_{\boldsymbol{Y}|\boldsymbol{N}}(\boldsymbol{y}_i|\boldsymbol{n}_i)\\&=\sum_{j=1}^{J_i}\ln f_{Y_t|N_t}(y_{it_{ij}}|n_{it_{ij}})+\ln c^{\mathrm{Sev}}_{T:t_{i1},\dots,t_{iJ_i}}(\omega_{i1},\dots,\omega_{iT}),\end{aligned}$$

where $\omega_{i1},\dots,\omega_{iT}$ are defined as Eq. (7). Note that $\hat{\boldsymbol{\Lambda}}_2$ which maximizes (19) can be found by maximizing $\sum_{i=1}^{m}\ell_i(\boldsymbol{\zeta}^{\mathrm{Freq}})$ and $\sum_{i=1}^{m}\ell_i(\boldsymbol{\zeta}^{\mathrm{Sev}},\boldsymbol{\zeta}^{\mathrm{FS}})$ separately.

The large sample properties of the IFM estimators for copula models are studied by Joe and Xu (1996) and Joe (2005). The more general treatment of the stage-wise estimators can be found in Newey and McFadden (1994). Let $\hat{\boldsymbol{\Lambda}} = \left(\hat{\boldsymbol{\Lambda}}_1', \hat{\boldsymbol{\Lambda}}_2'\right)'$ denote the IFM estimator. Define the inference function vector as

$$\boldsymbol{S}(\boldsymbol{\Lambda})=\begin{pmatrix}\partial\ell(\boldsymbol{\theta}^{\mathrm{Freq}})/\partial\boldsymbol{\theta}^{\mathrm{Freq}}\\ \partial\ell(\boldsymbol{\theta}^{\mathrm{Sev}})/\partial\boldsymbol{\theta}^{\mathrm{Sev}}\\ \partial\ell(\boldsymbol{\zeta}^{\mathrm{Freq}})/\partial\boldsymbol{\zeta}^{\mathrm{Freq}}\\ \partial\ell(\boldsymbol{\zeta}^{\mathrm{Sev}},\boldsymbol{\zeta}^{\mathrm{FS}})/\partial\boldsymbol{\zeta}^{\mathrm{Sev}}\\ \partial\ell(\boldsymbol{\zeta}^{\mathrm{Sev}},\boldsymbol{\zeta}^{\mathrm{FS}})/\partial\boldsymbol{\zeta}^{\mathrm{FS}}\end{pmatrix}$$

Under some regularity conditions, the IFM estimator is consistent and asymptotically normally distributed. i.e., $m\to\infty$,

$$\sqrt{m}(\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda})\xrightarrow{d}N(0,\boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega}$ is the inverse Godambe information matrix (Godambe, 1960) defined as $\boldsymbol{\Omega}=\boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}$ with $\boldsymbol{H}=-\mathrm{E}\left[\partial\boldsymbol{S}(\boldsymbol{\Lambda})/\partial\boldsymbol{\Lambda}'\right]$ and $\boldsymbol{V}=\mathrm{E}\left[\boldsymbol{S}(\boldsymbol{\Lambda})\boldsymbol{S}'(\boldsymbol{\Lambda})\right]$. Thus the asymptotic covariance matrix for $\hat{\boldsymbol{\Lambda}}$

can be approximated by $\hat{\Omega}^{-1}/m$, where the sample estimates of $\boldsymbol{H}$ and $\boldsymbol{V}$ are obtained, respectively, by

$$\hat{\boldsymbol{H}} = -\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial \boldsymbol{\Lambda}'}\boldsymbol{S}_i(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\hat{\boldsymbol{\Lambda}}} \quad \text{and} \quad \hat{\boldsymbol{V}} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{S}_i(\boldsymbol{\Lambda})\boldsymbol{S}_i'(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\hat{\boldsymbol{\Lambda}}}.$$

We note that the IFM method enjoys computational gain at the price of efficiency loss, which is arguably less critical for predictive applications. If the statistical efficiency is of primary interest, one common strategy is to perform a full maximum likelihood estimation, where one uses the estimates from the IFM method as initial values when maximizing the full likelihood function.

4.2. Prediction

The type of statistical inference that is of particular importance to our application is prediction. This section shows the predictive distribution of variables of interest to insurers. The term "predictive distribution" is used in a similar sense as in a Bayesian context, referring to the conditional distribution of the future outcome given past outcomes. It is, however, worth noting that the predictive distribution derived from the proposed model differs from the term that is mostly used in a Bayesian context. In the Bayesian framework, one treats parameters as random variables and uses the data to update the distribution of parameters which further serves as the mixing weight in the resulting predictive distribution. In contrast, our predictive distribution does not include uncertainty relating to model parameters which are treated as fixed quantities.

4.2.1. General Case

From the proposed model, one could derive the predictive distributions for claim frequency, claim severity, and the loss cost at the policyholder level. Statistics associated with these predictive distributions are often the key input for insurance operations such as underwriting, ratemaking, and claims management, among others. Note that the predictive distributions presented below are conditional on covariates which are suppressed to simplify notations.

The predictive distribution of frequency N_{iT+1} given $\boldsymbol{N}_i$ can be written as

$$\begin{aligned} &F_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i) \\ &= \Pr(N_{iT+1} \le n|\boldsymbol{N}_i = \boldsymbol{n}_i) \\ &= \frac{\sum_{l_1=0}^{1}\cdots\sum_{l_T=0}^{1}(-1)^{l_1+\cdots+l_T} \times C_{T+1}^{\text{Freq}}(u_{i1,l_1},\ldots,u_{iT,l_T},F_{N_{iT+1}}(n))}{\sum_{l_1=0}^{1}\cdots\sum_{l_T=0}^{1}(-1)^{l_1+\cdots+l_T} \times C_{T}^{\text{Freq}}(u_{i1,l_1},\ldots,u_{iT,l_T})}. \end{aligned} \tag{20}$$

And the associated probability mass function is calculated as

$$\begin{aligned} f_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i) &= \Pr(N_{iT+1} = n|\boldsymbol{N}_i = \boldsymbol{n}_i) \\ &= F_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i) - F_{N_{iT+1}|\boldsymbol{N}_i}(n-1|\boldsymbol{n}_i). \end{aligned} \tag{21}$$

Given the number of claims, we can discuss the distribution of the average severity. Specifically, the predictive distribution of Y_{iT+1} given $N_{iT+1} = n_{iT+1} > 0$, $\boldsymbol{Y}_i$, and $\boldsymbol{N}_i$ is

$$\begin{aligned} &F_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}(y|n_{iT+1},\boldsymbol{y}_i,\boldsymbol{n}_i) \\ &= \Pr(Y_{iT+1} \le y|N_{iT+1} = n_{iT+1},\boldsymbol{Y}_i = \boldsymbol{y}_i,\boldsymbol{N}_i = \boldsymbol{n}_i) \\ &= \frac{c^{\text{Sev}}_{T+1:t_{i1},\ldots,t_{iJ_i}}\left(\omega_{i1},\ldots,\omega_{iT},F_{Y_{iT+1}|N_{iT+1}}(y|n_{iT+1})\right)}{c^{\text{Sev}}_{T:t_{i1},\ldots,t_{iJ_i}}\left(\omega_{i1},\ldots,\omega_{iT}\right)}. \end{aligned} \tag{22}$$

The associated predictive density is

$$\begin{aligned} &f_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}(y|n_{iT+1},\boldsymbol{y}_i,\boldsymbol{n}_i) \\ &= \frac{\partial}{\partial y}F_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}(y|n_{iT+1},\boldsymbol{y}_i,\boldsymbol{n}_i) \\ &= f_{Y_{T+1}|N_{T+1}}(y|n_{iT+1}) \\ &\quad \times \frac{c^{\text{Sev}}_{T+1:t_{i1},\ldots,t_{iJ_i},T+1}\left(\omega_{i1},\ldots,\omega_{iT},F_{Y_{iT+1}|N_{iT+1}}(y|n_{iT+1})\right)}{c^{\text{Sev}}_{T:t_{i1},\ldots,t_{iJ_i}}\left(\omega_{i1},\ldots,\omega_{iT}\right)}. \end{aligned} \tag{23}$$

For $N_{iT+1} = n_{iT+1} = 0$, $Y_{iT+1} = 0$ with probability one.

Given the above predictive distributions, one could easily calculate the risk scores that are relevant to decision making in insurance operations. For instance, the expected claim frequency can be obtained by

$$\begin{aligned} \hat{N}_{iT+1} = E[N_{iT+1}|\boldsymbol{N}_i] &= \sum_{n=0}^{\infty} n f_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i) \\ &\approx \sum_{n=0}^{N_{max}} n f_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i) \end{aligned} \tag{24}$$

where N_{max} is a predefined large number. The expected loss cost can be calculated by

$$\begin{aligned} \hat{S}_{iT+1} &= E[N_{iT+1}Y_{iT+1}|\boldsymbol{N}_i,\boldsymbol{Y}_i] \\ &= \sum_{n=0}^{\infty}\left[n f_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i)\int_{y=0}^{\infty} y f_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}(y|n,\boldsymbol{y}_i,\boldsymbol{n}_i)dy\right] \\ &\approx \sum_{n=0}^{N_{max}}\left[n f_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i)\int_{y=0}^{\infty} y f_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}(y|n,\boldsymbol{y}_i,\boldsymbol{n}_i)dy\right]. \end{aligned} \tag{25}$$

An alternative approach to calculating the predictive distribution and the associated summary statistics is to perform a simulation. The frequency and severity components in the proposed model can be simulated in a sequential manner. Below we summarize the procedure for policyholder i. Let superscript (k) be the simulation index to indicate outcomes generated from the kth iteration. For $k = 1,\ldots,K$, repeat the following three steps so that one has a random sample of N_{iT+1} and S_{iT+1}:

1. Generate the number of claims $n_{iT+1}^{(k)}$ from the predictive distribution $f_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i)$ using the standard inversion method:

$$n_{iT+1}^{(k)} = F^{-1}_{N_{iT+1}|\boldsymbol{N}_i}\left(u_i^{(k)}|\boldsymbol{n}_i\right),$$

where $F^{-1}(q) = \inf\{x \in R : F(x) \ge q\}$, and $u_i^{(k)}$ is a random number on uniform (0, 1).

2. If $n_{iT+1}^{(k)} = 0$, then set $y_{iT+1}^{(k)} = 0$. Otherwise, generate average claim amount $y_{iT+1}^{(k)}$ from the predictive distribution $f_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}(y|n_{iT+1},\boldsymbol{y}_i,\boldsymbol{n}_i)$ using

$$y_{iT+1}^{(k)} = F^{-1}_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}\left(v_i^{(k)}|n_{iT+1}^{(k)},\boldsymbol{y}_i,\boldsymbol{n}_i\right),$$

where $v_i^{(k)}$ is a random number on uniform (0, 1).

3. Calculate loss cost $s^{(k)} = n_{iT+1}^{(k)} \times y_{iT+1}^{(k)}$.

For a large K, we can consistently estimate the predictive distribution for the frequency and loss cost, and the associated risk scores.

4.2.2. Special case

In our application, the special structure of AR1 dependence suggests only using the claim history of the last year in the prediction. It is straightforward to show that the predictive density for severity (23) reduces to

$$\begin{aligned} &f_{Y_{iT+1}|N_{iT+1},\boldsymbol{Y}_i,\boldsymbol{N}_i}(y|n_{iT+1},\boldsymbol{y}_i,\boldsymbol{n}_i) \\ &= f_{Y_{T+1}|N_{T+1}}(y|n_{iT+1}) \cdot c^{\text{Sev}}\left(F_{Y_{iT+1-\tau}|N_{iT+1}}(y_{iT+1-\tau}|n_{iT+1-\tau}),\right. \\ &\qquad \left. F_{Y_{iT+1}|N_{iT+1}}(y|n_{iT+1});\sigma^{\tau}_{sev}\right) \end{aligned} \tag{26}$$

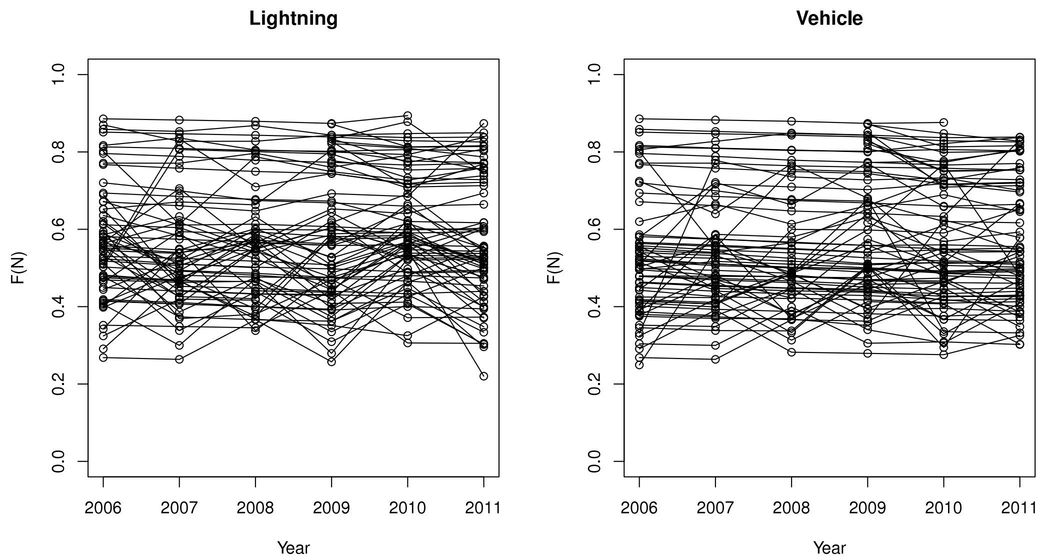

Fig. 2. Multiple timeseries plot of the Cox–Snell residuals from the marginal frequency model by peril type.

where $T+1-\tau$ corresponds to the most recent year with positive claim amount, and $c^{\text{Sev}}(\cdot;\sigma^{\tau}_{sev})$ is a bivariate Gaussian copula density with association parameter σ^{τ}_{sev}.

Unlike the severity, prediction for frequency does not lead to the above simplification. However, the AR1 dependence suggests that the most recent claim history is more predictive for the future. To improve the computational efficiency in the prediction, one could use the number of claims in the most recent years instead of the entire claim history in the frequency prediction. For instance, if one uses the claim history in the recent κ years in the prediction, the predictive distribution function for frequency (22) becomes Eq. (27) which is given in Box III. The above approximation could improve the computational efficiency significantly when the number of years of claim history T is large. We emphasize that the prediction of loss cost is computationally expensive even for moderate T because of the infinite sum in (25) together with the integral. We investigate the trade-off between computational efficiency and predictive precision in our empirical study.

5. Empirical analysis

5.1. Estimation results

We use the data in years 2006–2010 to develop the model, and reserve the data of year 2011 for hold-out sample validation. Tables 8 and 9 report the estimation results for the marginal models of the frequency and severity components, respectively. To recap, we have used the zero–one inflated negative binomial (NB) model for the frequency, and the *GB2* regression for the severity. According to Table 8, cities, counties and villages have high lightning claim frequencies, and also AC10 seems to be a significant indicator for high lightning frequencies. The coverage amount is also positively correlated with the lightning frequencies. For vehicle frequencies, it is notable that school entities had significantly lower vehicle claims than other entities. Also it appears that AC15 is positively related with vehicle frequencies. In addition, the coefficient associated with lnDeductBC is negative for both the lightning and vehicle marginal models. In contrast, the effect of deductible on severity is positive for both lightning and vehicle perils, although it is only statistically significant for the lightning peril. In the severity model for both perils, lnCoverageBC turns out to have a positive and significant effect, indicating larger entities are more likely to have higher claims.

Table 8
Estimation results for the zero–one inflated NB model by peril type.

	Lightning			Vehicle	
	Coef.	t-ratio		Coef.	t-ratio
NB regression			NB regression		
Intercept	−1.356	−3.010	Intercept	−2.777	−5.209
Type:City	0.753	2.973	Type:City	0.614	1.925
Type:County	1.342	5.148	Type:County	−0.588	−1.661
Type:School	−0.051	−0.200	Type:School	− 2.618	−7.166
Type:Town	0.306	0.879	Type:Town	0.550	1.047
Type:Village	0.805	3.174	Type:Village	0.470	1.397
AC05	−0.031	−0.118	AC05	0.185	0.429
AC10	0.381	2.065	AC10	0.190	0.550
AC15	0.168	1.596	AC15	0.388	2.296
lnDeductBC	−0.294	−4.386	lnDeductBC	−0.217	−3.091
lnCoverageBC	0.466	6.272	lnCoverageBC	0.903	8.428
HighFreq	0.690	1.946	HighFreq	1.401	3.101
Size	3.715	1.255	Size	0.706	4.011
Zero model			Zero model		
Intercept	−1.975	−1.977	Intercept	−2.604	−2.622
lnDeductBC	0.430	3.275	lnDeductBC	0.660	5.259
lnCoverageBC	−0.539	−3.952	lnCoverageBC	−0.678	−4.547
One model			One model		
Intercept	79.628	0.064	Intercept	5.456	0.631
lnDeductBC	−15.570	−0.078	lnDeductBC	−1.297	−0.924
lnCoverageBC	2.730	1.100	lnCoverageBC	−0.399	−1.844

Fig. 2 displays the multiple time series plots of a random sample of the Cox–Snell residuals from the marginal frequency model for both lightning and vehicle perils. The Cox–Snell residuals, defined as $\widehat{F}_{N_t}(n_{it})$, are calculated using the estimated parameters of the claim frequency distribution which are reported in Table 8. After removing the effects of explanatory variables, the time series plots suggest moderate unobserved heterogeneity over time. In addition, the claim frequency for the vehicle peril exhibits higher subject-specific effects than the lightning peril. These patterns are supported by the serial correlations reported in Tables 5 and 6.

To assess the goodness-of-fit of the fitted models, we compare the observed data with the fitted observations using the estimated model. For the frequency component, Table 10 reports the

$$F_{N_{iT+1}|\boldsymbol{N}_i}(n|\boldsymbol{n}_i) \approx \frac{\sum_{l_{T+1-\kappa}=0}^{1}\cdots\sum_{l_{T+1}=0}^{1}(-1)^{l_{T+1-\kappa}+\cdots+l_{T+1}} \times C_{T+1}^{\text{Freq}}(u_{iT+1-\kappa,l_{T+1-\kappa}},\ldots,u_{iT,l_T},F_{N_{iT+1}}(n))}{\sum_{l_{T+1-\kappa}=0}^{1}\cdots\sum_{l_{T}=0}^{1}(-1)^{l_{T+1-\kappa}+\cdots+l_{T}} \times C_{T}^{\text{Freq}}(u_{i1,l_1},\ldots,u_{iT,l_T})}. \tag{27}$$

Box III.

Table 9
Estimation results for the GB2 regression by peril type.

	Lightning			Vehicle	
	Coef.	t-ratio		Coef.	t-ratio
(Intercept)	6.552	12.466	(Intercept)	7.250	22.309
Type:City	−0.085	−0.325	Type:City	0.058	0.285
Type:County	0.200	0.734	Type:County	−0.213	−0.928
Type:School	0.301	1.123	Type:School	−0.268	−1.072
Type:Town	−0.195	−0.592	Type:Town	0.406	1.494
Type:Village	0.279	1.110	Type:Village	0.015	0.075
AC05	0.183	0.708	AC05	0.136	0.441
AC10	0.392	1.998	AC10	0.183	0.759
AC15	−0.025	−0.232	AC15	−0.019	−0.169
lnDeductBC	0.093	2.398	lnDeductBC	0.007	0.140
lnCoverageBC	0.159	3.371	lnCoverageBC	0.099	2.144
σ	1.032	2.794	σ	0.516	2.901
α_1	3.385	1.422	α_1	1.421	1.596
α_2	1.968	1.780	α_2	0.949	2.241

Table 10
Empirical and fitted claim frequency by peril type.

Claim count	Lightning		Claim count	Vehicle	
	Empirical	Fitted		Empirical	Fitted
0	5044	5041.824	0	5250	5255.655
1	461	475.967	1	265	251.511
2	112	100.324	2	62	67.037
3	29	26.735	3	27	30.712
4	11	8.823	4	23	16.745
5	2	3.402	5	8	10.135
6	1	1.470	6	5	6.592

fitted number of policyholders along with the observed number of policyholders by claim count. To account for observed heterogeneity, the fitted number of policyholders is calculated by summing up the fitted probability of a given number of events over all policyholders. The results suggest that the zero–one inflated NB model is a reasonable fit. In addition, we also explored commonly used count regression models including Poisson and negative binomial, as well as the related zero-inflated models. The chi-square statistics supports the selected zero–one inflated NB model. For brevity, we did not report the goodness-of-fit results for all candidate models.

For the severity, we examine the Cox–Snell residuals that remove the effect of covariates. The Cox–Snell residual is defined by $\widehat{F}_{GB2}(y_{it}|N_{it} > 0)$ where $\widehat{F}_{GB2}$ is the fitted $GB2$ regression. Fig. 3 reports the histogram of the residuals, where the uniformity suggests the good fit of the model. In addition, we present in Fig. 3 the normal QQ-plots of the residuals. Specifically, we transform the Cox–Snell residuals to normal scores using $\Phi^{-1}\left(\widehat{F}_{GB2}(y_{it}|N_{it} > 0)\right)$. In agreement with the histogram, the QQ-plots show the consistency between the empirical quantiles and the fitted quantiles.

Table 11 reports the estimated association parameters in the copula models, i.e. the temporal dependence in frequency, the temporal dependence in severity, as well as the contemporaneous dependence between the frequency and severity. Results

Table 11
Estimated association parameters in the copula model by peril type.

	Lightning			Vehicle	
	Coef.	Std.Err.		Coef.	Std.Err.
ρ_{FS}	−0.023	0.055	ρ_{FS}	−0.167	0.064
ρ_{Freq}	0.322	0.068	ρ_{Freq}	0.621	0.044
ρ_{Sev}	0.202	0.157	ρ_{Sev}	0.378	0.156

from the dependence modeling indicate that high serial correlations exist after controlling for the explanatory variables using marginal models. In general, the estimated results are consistent with the preliminary analysis reported in Section 2. For the lightning peril, the frequency exhibits significant positive serial dependence. The serial correlation in severity given frequency is moderate although not statistically significant. The dependence between claim frequency and average severity is insignificant, as we have expected from the weak polyserial correlation.

The vehicle claims show a negative and significant dependence between the claim frequency and average severity. Also, the high serial dependence for both the frequency and severity is notable. Our estimation results suggest that the prediction results may improve when the longitudinal nature of recurrent insurance claims by peril type is exploited.

As discussed in Section 3.2.2, there are strengths and limitations with the Gaussian copula. For our application, the Gaussian copula is employed mainly to balance the interpretability, flexibility, as well as the tractability of the model. Specifically, the Gaussian copula is simple and easy to understand for an applied audience. And more importantly, the dispersion matrix with AR(1) structure has a natural interpretation in the longitudinal context. To demonstrate that the Gaussian copula with AR(1) dependence is appropriate for the real data, we display in Fig. 4 the pair-wise pp-plots by time lag for the severity model. The pair-wise plot is motivated by the unbalanced nature of the conditional severity data. Specifically we examine the empirical cdf of the bivariate copula and the theoretical cdf of the Gaussian copula with parameter $\hat{\rho}^k$ where $\hat{\rho}$ is the estimated association parameter and $k = 1, 2, 3, 4$ denotes the time lag. The result suggests that the Gaussian copula with AR(1) dependence is supported by the data.

5.2. Out-of-sample validation

Recall that the proposed copula model is estimated using data from year 2006 to 2010. For validation purposes, we use the fitted model to make predictions for the loss cost in year 2011, and compare the predicted loss cost with the actual observed loss cost in the hold-out sample. Tables 12 and 13 summarize the correlation coefficients between the actual and predicted loss costs for lightning and vehicle perils, respectively. The predicted loss costs are calculated using (25), where a policyholder's entire claim history is used in the prediction of future outcome. For comparison, we also report the results for predictions using the independence model, where neither the temporal correlation in the frequency, the temporal correlation in the severity, nor the

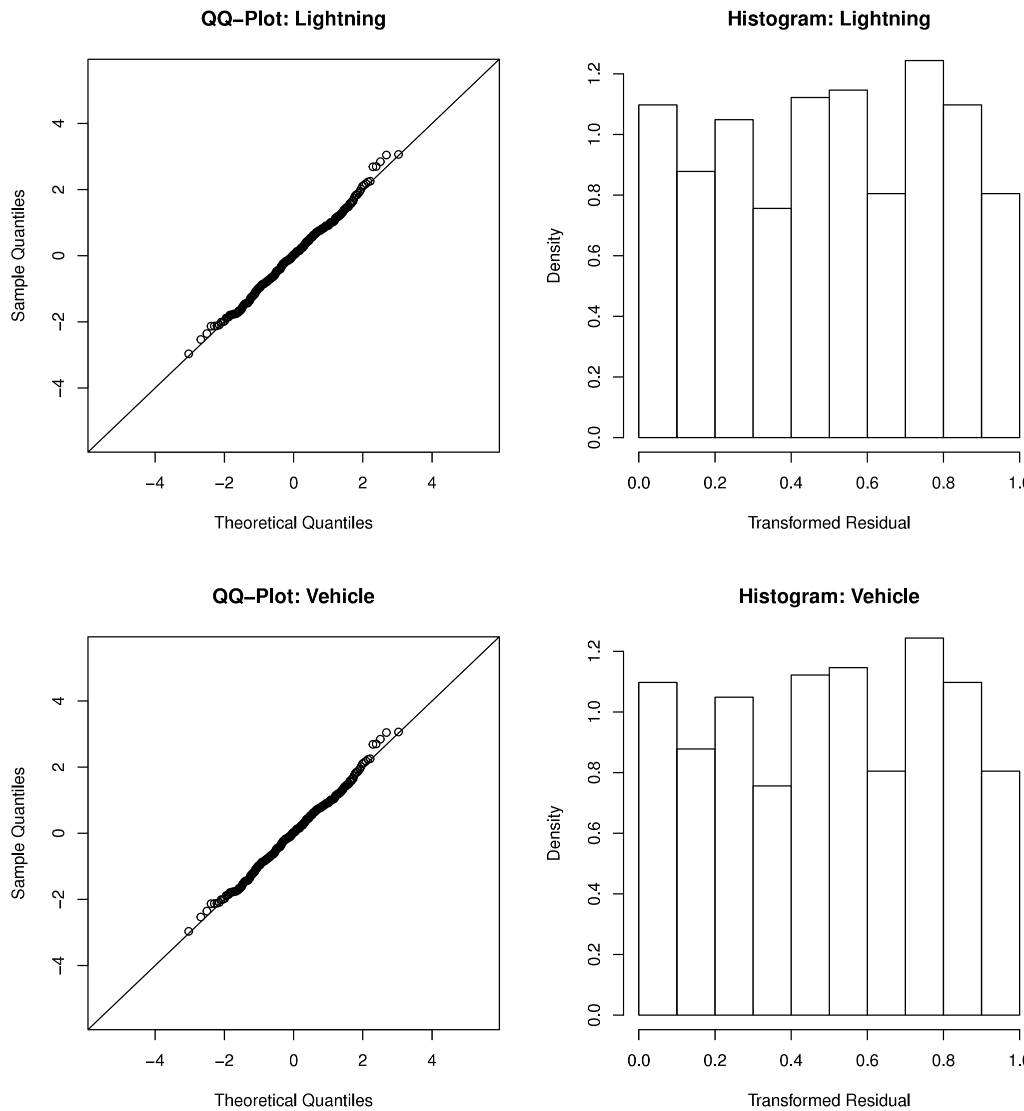

Fig. 3. The QQ-Plots of normal scores and the histograms of the Cox–Snell residuals from the marginal severity model by peril type.

Table 12
Correlation between actual and predicted loss cost for lightning peril.

	Pearson correlation	Spearman correlation
Independence model	38.053	27.884
Dependence model	38.494	29.689

Table 13
Correlation between actual and predicted loss cost for vehicle peril.

	Pearson correlation	Spearman correlation
Independence model	32.138	34.765
Dependence model	50.181	35.601

contemporaneous correlation between the frequency and severity is considered. The results suggest an improvement in the prediction using the longitudinal model. The higher lift in the vehicle peril is consistent with the strong relation for both within and between frequency and severity components. Recall that for the lightning peril, only the temporal relation within the frequency was found to be significant.

Because of the large portion of zeros in the loss costs, the above reported correlation coefficients are sometimes not informative. We provide additional validation tests by looking into the ordered Lorenz curve and the associated Gini index for both the claim frequency and the loss cost. The concept of using the ordered Lorenz curve and Gini index to select insurance risk is introduced in Frees et al. (2011b). The central idea of the ordered Lorenz curve is to compare the premium distribution and the loss distribution that are both ordered by a relativity. The relationship between the premium and loss distribution informs us whether the defined relativity could facilitate profitable risk selection. This relationship is summarized by the Gini index — twice the area between the ordered Lorenz curve and the line of equality, which is the 45 degree line. Mathematically, let $G_P(r)$ and $G_L(r)$ denote the premium and loss distributions, respectively, the Gini index can be written as

$$\mathrm{Gini}(G_P, G_L) = 2\int_0^\infty (G_P(r) - G_L(r))dG_P(r),$$

which can be shown to take a value in $[-1, 1]$.

We use this approach to evaluate the predictions for both claim frequency and loss cost. The predicted claim frequencies

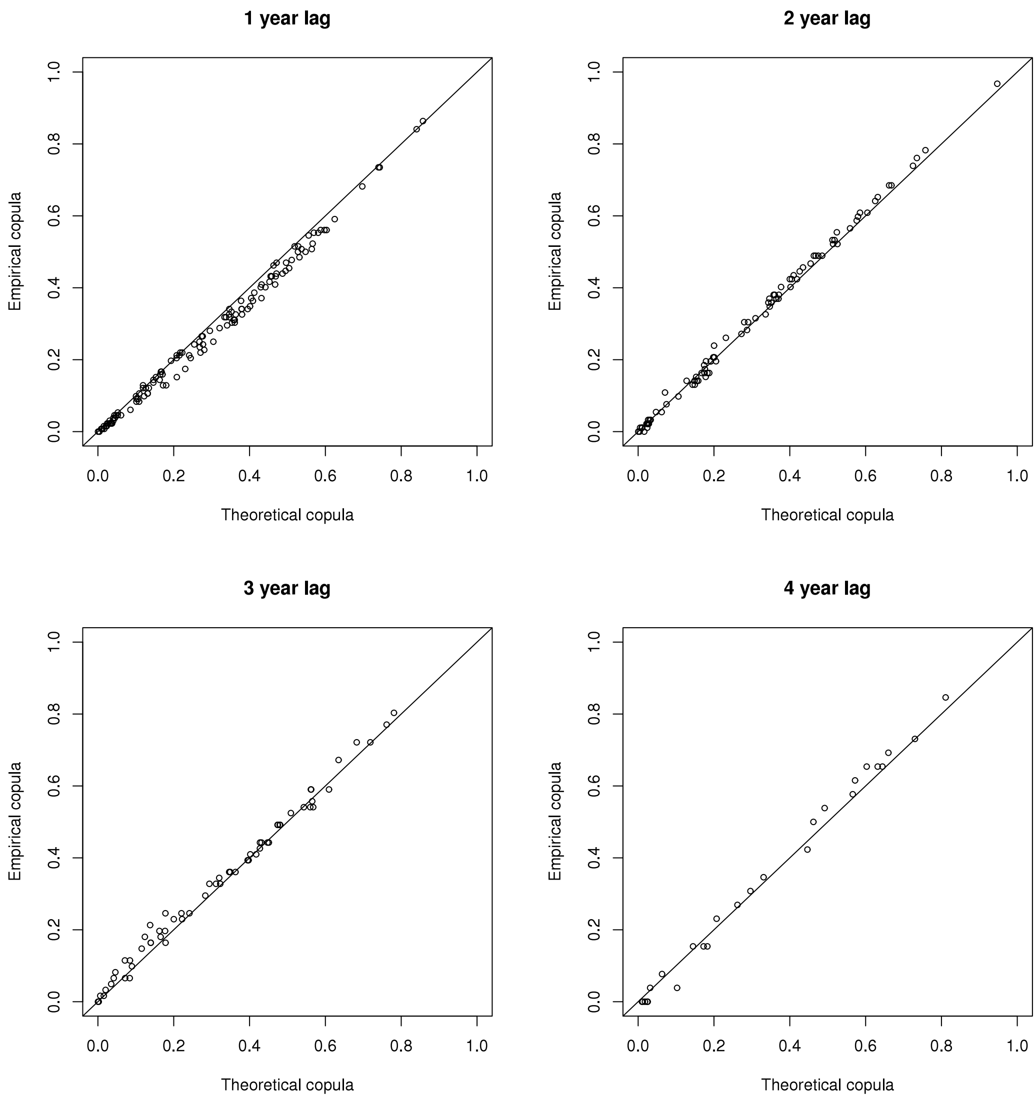

Fig. 4. The pair-wise pp-plots for the Gaussian copula by time lag.

and loss costs are calculated using (24) and (25), respectively, for the copula model. The prediction for the independence model is calculated using marginal distributions, i.e. the zero–one inflated NB regression for frequency and the GB2 regression for severity. For comparison, we use the prediction from the independence model as the base premium, and use the prediction from the copula model as the competing premium. The relativity is defined as the ratio of the competing premium to the base premium. Fig. 5 displays the ordered Lorenz curve for the claim frequency and the loss cost by peril types. The results suggest that the insurer could identify more profitable portfolios by looking at the premium implied by the copula model.

The corresponding Gini indices and standard errors are reported in Table 14. The positive indices and statistical significance reinforce the conclusions that are drawn from the ordered Lorenz curve. In addition, the Gini index for the vehicle peril is larger than the lightning peril. This result is consistent with the stronger dependence that we have observed for the vehicle peril.

5.3. Prediction comparison

This section investigates the performance of the prediction using a policyholder's claim experience in the most recent year as opposed to the entire claim history. This approximation is motivated by the AR1 serial dependence, which suggests that observations further apart in time are less correlated. We compare the approximate prediction with the exact prediction, and report the results for the vehicle peril as an illustration.

Table 14
Gini indices and standard errors for claim frequency and loss cost by peril type.

	Frequency			Loss cost	
	Lightning	Vehicle		Lightning	Vehicle
Gini index	0.242	0.473	Gini index	0.132	0.331
Standard error	(0.060)	(0.054)	Standard error	(0.082)	(0.101)

The first experiment focuses on the claim frequency. We predict the number of claims for policyholders in the hold-out-sample using different number of years of claim history. The scatter plot matrix is displayed in Fig. 6 to visualize the comparison. Zero claim history means using the marginal model in the prediction and ignoring the serial dependence in the claim frequency. The scatter plot comparing predictions of the independence model and the one-year claim history model suggests the

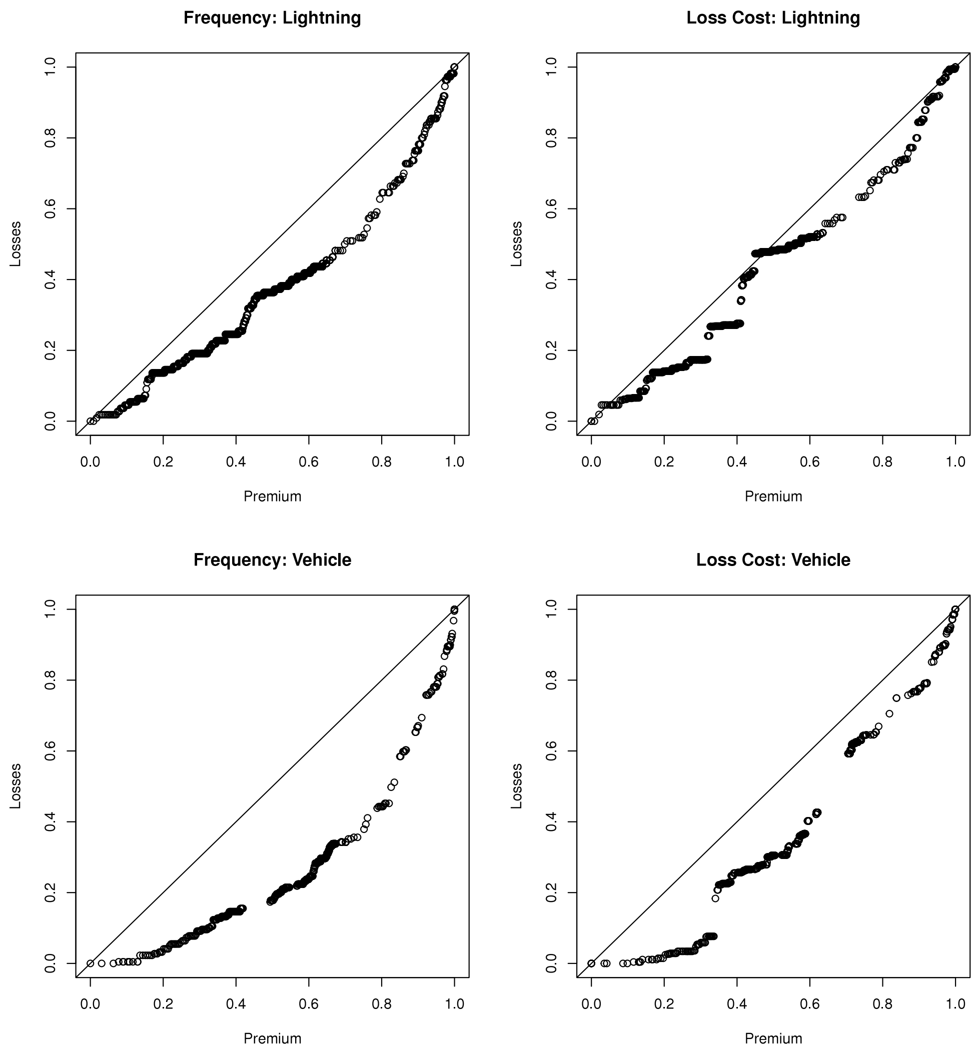

Fig. 5. Ordered Lorenz curves for claim frequency and loss cost by peril type.

Table 15
Comparison of validation statistics for loss cost predictions.

	1-year history	5-year history
Pearson correlation	50.181	50.333
Spearman correlation	35.601	35.700
Gini index	0.331	0.313
(Standard error)	(0.101)	(0.116)

critical role of the temporal relationship among claim frequency in the prediction. The result also shows that using additional years of claim history does not provide much lift in the frequency prediction.

The second experiment focuses on the loss cost. For the policyholders in the hold-out sample, we predict their loss costs using the claim history of the most recent year and using the entire claim history. To compare the predictions, we refer to the out-of-sample validation statistics as in Tables 13 and 14. Specifically we calculate the correlation between predicted loss costs and the actual loss costs, and calculate the Gini indices using the prediction from the independence model as the base premium. The validation statistics are reported in Table 15. The results confirm that the claim history in recent years matter most for the prediction.

6. Concluding remarks

Understanding the factors that contribute to repeated insurance claims can help to mitigate the risk of future insurance claims. In this paper, we have provided a framework for modeling recurrent insurance claims in a longitudinal setup using copulas to capture the dependence of claim frequencies over time, the dependence of average claim severities over time, as well as the dependence between the frequencies and severities. Through an empirical study utilizing the LGPIF data, we attempted to explain the factors that contribute to frequent insurance claim peril types such as lightning and vehicle. The marginal models indicate that entity type, deductible level, and coverage amount are significant predictors of the perils. Yet, even after controlling for the explanatory variables, we were able to discover that there exists serial dependence of the claim frequencies, and severities over time. By exploiting these dependencies and constructing a model, which captures this dependence, we were able to improve the claim score prediction for the frequent peril types.

Our model shows that the prediction results clearly improve. Although a promising approach, the method is in its infancy, and still needs improvements. One important improvement needed is the computational time required for the prediction. For applications to large datasets, the computational time can be reduced

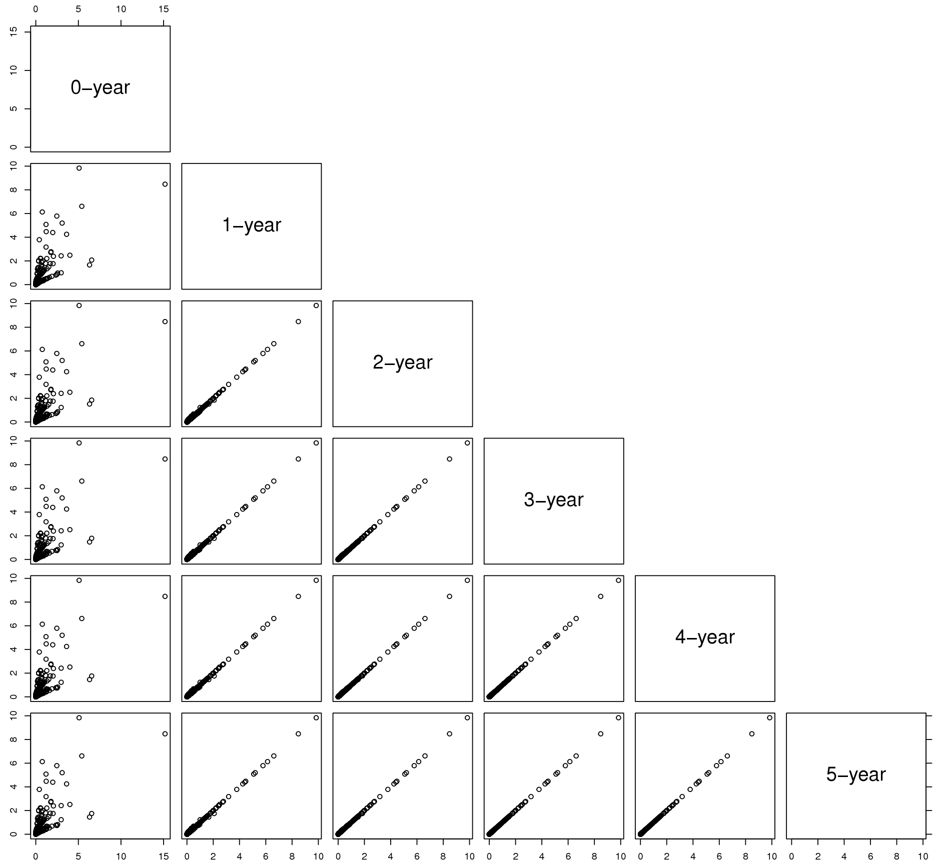

Fig. 6. Scatter plot matrix of predicted frequency using 0–5 years of claim history.

by utilizing parallel processing. With more computational power, more historic information can be incorporated into the model, and hence better prediction results will be achievable.

Because our approach allows for accurate prediction of both the frequencies and the claim scores of insurance claims, we believe that wide applications will be possible. In our work, the normal copula has been used to capture the dependence structure over time. This allows for flexible modeling of the dependence structures, as well as estimation. Our approach provides a simple and flexible approach to modeling the dependence structure of non-normal response variables. The philosophy of our modeling approach can be applied to other non-normal response variables, such as ordinal variables. In our work, we have focused on modeling the building and contents coverage group, although other coverage groups with complex dependence structures exist in the dataset. See Frees et al. (2016). Future work may focus on incorporating the dependencies among lines, while considering the longitudinal nature of various peril types occurring to coverages with complex dependence structures.

Acknowledgment

We thank the editor and anonymous reviewers for their valuable comments that helped improve the paper significantly. Peng Shi acknowledges the support from the Society of Actuaries CAE Research grant.

References

Boucher, J.-P., 2014. Regression with count dependent variables. In: Frees, E.W., Meyers, G., Derrig, R.A. (Eds.), Predictive Modeling Applications in Actuarial Science. Cambridge University Press, Cambridge.
Boucher, J.-P., Denuit, M., Guillén, M., 2008. Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. Variance 2 (1), 135–162.
Boucher, J.-P., Guillén, M., 2011. A semi-nonparametric approach to model panel count data. Comm. Statist. Theory Methods 40 (4), 622–634.
Cook, M.A., 2012. Survey of Personal Insurance and Financial Planning. The Institutes.
Czado, C., Kastenmeier, R., Brechmann, E.C., Min, A., 2012. A mixed copula model for insurance claims and claim sizes. Scand. Actuar. J. 2012 (4), 278–305.
Erhardt, V., Czado, C., 2012. Modeling dependent yearly claim totals including zero claims in private health insurance. Scand. Actuar. J. 2012 (2), 106–129.
Flitner, A.L., 2014. Survey of Commercial Insurance, second ed. The Institutes.
Frees, E., 2014. Frequency and severity models. In: Frees, E., Meyers, G., Derrig, R.A. (Eds.), Predictive Modeling Applications in Actuarial Science, Volume I: Predictive Modeling Techniques. Cambridge University Press, Cambridge, pp. 138–166.
Frees, E.W., Gao, J., Rosenberg, M.A., 2011a. Predicting the frequency and amount of health care expenditures. N. Am. Actuar. J. 15 (3), 377–392.
Frees, E.W., Lee, G.Y., Yang, L., 2016. Multivariate frequency-severity regression models in insurance. Risks 2016 (4).
Frees, E.W., Meyers, G., Cummings, D., 2010. Dependent multi-peril ratemaking models. Astin Bull. 699–726.
Frees, E.W., Meyers, G., Cummings, A.D., 2011b. Summarizing insurance scores using a Gini index. J. Amer. Statist. Assoc. 106 (495).
Garrido, J., Genest, C., Schulz, J., 2016. Generalized linear models for dependent frequency and severity of insurance claims. Insurance Math. Econom. 70, 205–215.
Godambe, V.P., 1960. An optimum property of regular maximum likelihood estimation. Ann. Math Stat. 31 (4), 1208–1211.
Gschlößl, S., Czado, C., 2007. Spatial modelling of claim frequency and claim size in non-life insurance. Scand. Actuar. J. 2007 (3), 202–225.
Hua, L., 2015. Tail negative dependence and its applications for aggregate loss modeling. Insurance Math. Econom. 61, 135–145.

Joe, H., 2005. Asymptotic efficiency of the two-stage estimation method for copula-based models. J. Multivariate Anal. 94 (2), 401–419.
Joe, H., 2014. Dependence Modeling with Copulas. CRC Press.
Joe, H., Xu, J.J., 1996. The Estimation Method of Inference Functions for Margins for Multivariate Models, Technical Report. UBC, Department of Statistics, p. 166.
Klugman, S.A., Panjer, H.H., Willmot, G.E., 2012. Loss Models: From Data to Decisions. John Wiley & Sons.
Krämer, N., Brechmann, E.C., Silvestrini, D., Czado, C., 2013. Total loss estimation using copula-based regression models. Insurance Math. Econom. 53 (3), 829–839.
Lee, G., 2017. General insurance deductible ratemaking. N. Am. Actuar. J. 21 (4), 620–638.
Nelsen, R.B., 1999. An Introduction to Copulas, Vol. 139. Springer Science & Business Media.
Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. Handb. Econom. 4, 2111–2245.
Panagiotelis, A., Czado, C., Joe, H., 2012. Pair copula constructions for multivariate discrete data. J. Amer. Statist. Assoc. 107 (499), 1063–1072.
Shi, P., 2014. Fat-tailed regression models. In: Frees, E.W., Meyers, G., Derrig, R.A. (Eds.), Predictive Modeling Applications in Actuarial Science. Cambridge University Press, Cambridge.
Shi, P., Feng, X., Ivantsova, A., 2015. Dependent frequency–severity modeling of insurance claims. Insurance Math. Econom. 64, 417–428.
Shi, P., Valdez, E.A., 2014. Longitudinal modeling of insurance claim counts using jitters. Scand. Actuar. J. 2014 (2), 159–179.