UNIVERSIDADE FEDERAL DA PARAÍBA
DEPARTAMENTO DE FINANÇAS E CONTABILIDADE
CURSO DE BACHARELADO EM CIÊNCIAS ATUARIAIS

MARCIO VAZ DE OLIVEIRA

A SINISTRALIDADE NO SEGURO DE RESPONSABILIDADE CIVIL GERAL

JOAO PESSOA
2026

MARCIO VAZ DE OLIVEIRA

A SINISTRALIDADE NO SEGURO DE RESPONSABILIDADE CIVIL GERAL

Monografia apresentada ao curso de Ciências Atuariais da Universidade Federal da Paraíba, como exigência para a conclusão do curso e obtenção do grau de bacharel.

Orientador: Prof. Dr. Filipe Coelho de Lima Duarte

JOÃO PESSOA
2026

Catalogação na publicação
Seção de Catalogação e Classificação

O48s Oliveira, Marcio Vaz de.
A sinistralidade no Seguro de Responsabilidade Civil Geral / Marcio Vaz de Oliveira. - João Pessoa, 2026.
89 f. : il.

Orientação: Filipe Coelho de Lima Duarte.
TCC (Graduação) - UFPB/CCSA.

1. Seguros de responsabilidade civil. 2. Séries temporais. 3. Modelos lineares generalizados. 4. Mercado segurador. I. Duarte, Filipe Coelho de Lima. II. Título.

UFPB/CCSA CDU 368

Elaborado por KATIANE DA CUNHA SOUZA - CRB-15/353

A SINISTRALIDADE NO SEGURO DE RESPONSABILIDADE CIVIL GERAL

Monografia apresentada ao curso de Ciências Atuariais da Universidade Federal da Paraíba, como exigência para a conclusão do curso e obtenção do grau de bacharel.

Aprovado em:

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br
FILIPE COELHO DE LIMA DUARTE
Data: 01/04/2026 08:43:38-0300
Verifique em https://validar.iti.gov.br

Filipe Coelho de Lima Duarte (Orientador)

Universidade Federal da Paraíba

Documento assinado digitalmente
gov.br
SAMARA LAUAR SANTOS
Data: 02/04/2026 00:57:50-0300
Verifique em https://validar.iti.gov.br

Samara Lauar Santos

Universidade Federal da Paraíba

Documento assinado digitalmente
gov.br
YURI MARTI SANTANA SANTOS
Data: 01/04/2026 18:33:42-0300
Verifique em https://validar.iti.gov.br

Yuri Marti Santana Santos

Universidade Federal da Paraíba

RESUMO

Este trabalho investiga quais fatores observáveis do mercado segurador explicam a dinâmica da sinistralidade no ramo 0351 – Seguro de Responsabilidade Civil Geral – no Brasil. A análise foi conduzida a partir de dados mensais reportados pelas seguradoras à Superintendência de Seguros Privados (SUSEP), abrangendo o período de novembro de 2016 a maio de 2025. Inicialmente, foi realizada modelagem de séries temporais por meio de modelo SARIMA, com o objetivo de identificar padrões temporais e sazonais na sinistralidade do mercado. Em seguida, foi estimado um modelo vetorial autorregressivo (VAR) para examinar a interação entre a sinistralidade e variáveis representativas da estrutura operacional das seguradoras. Complementarmente, foram estimados modelos lineares generalizados (GLM) sob diferentes distribuições da família exponencial, sendo selecionado o modelo com distribuição Tweedie com base no critério de informação AIC. Também foi aplicada técnica de agrupamento para identificar segmentos distintos entre as seguradoras do mercado. Os resultados indicam que a sinistralidade apresenta persistência temporal e está associada a variáveis como retenção de riscos, participação em provisões técnicas e participação no prêmio do mercado. Conclui-se que a dinâmica da sinistralidade no ramo analisado pode ser explicada por fatores observáveis relacionados à estrutura operacional das seguradoras e às condições do mercado segurador.

Palavras-chave: sinistralidade; seguros de responsabilidade civil; séries temporais; modelos lineares generalizados; mercado segurador.

ABSTRACT

This study investigates which observable factors in the insurance market explain the dynamics of loss ratios in line 0351 – General Liability Insurance – in Brazil. The empirical analysis is based on monthly data reported by insurance companies to the Brazilian insurance regulator (SUSEP), covering the period from November 2016 to May 2025. Initially, a SARIMA time series model was estimated in order to identify temporal and seasonal patterns in market loss ratios. Subsequently, a vector autoregressive (VAR) model was applied to examine the interaction between loss ratios and variables representing the operational structure of insurance companies. In addition, generalized linear models (GLM) were estimated under different distributions from the exponential family. The Tweedie distribution provided the best fit according to the Akaike Information Criterion (AIC). A clustering algorithm was also employed to identify distinct segments among insurers operating in the market. The results indicate that loss ratios exhibit temporal persistence and are associated with variables such as risk retention, participation in technical reserves, and market premium share. The findings suggest that the dynamics of loss ratios in the analyzed line can be explained by observable factors related to insurers' operational structure and market conditions

Keywords: loss ratio; liability insurance; time series; generalized linear models; insurance market.

SUMÁRIO

1 INTRODUÇÃO

O seguro representa uma das atividades econômicas mais antigas do mundo e está disciplinado no Código Civil Brasileiro, como sendo o contrato pelo qual o segurador se obriga, mediante pagamento de um prêmio, a garantir interesse legítimo do segurado contra riscos predeterminados (BRASIL, 2002).

O contrato de seguro trata-se de um instrumento jurídico de socialização de riscos (Coelho, 2002), operacionalizado por meio de uma relação que pressupõe o equilíbrio entre o valor do prêmio pago e a garantia assumida. Essa relação securitária exige que o risco coberto seja possível e incerto (Alvim, 1999), além de lícito e economicamente mensurável, de modo a constituir interesse legítimo a ser protegido pela apólice.

Nesse contexto, determinadas atividades profissionais e empresariais estão expostas ao risco de responsabilização civil de seus agentes. Os danos civis, seja de natureza patrimonial ou moral, podem gerar elevada exposição financeira, fazendo surgir a demanda por instrumentos que mitiguem esse risco, sendo um deles o seguro de responsabilidade civil.

A responsabilidade civil, está fundamentada no artigo 927 do Código Civil Brasileiro, que, combinado aos artigos 186 e 187 do mesmo diploma, permitem conceituá-la, segundo Pereira (2004), como a obrigação de reparar o dano imposta a todo aquele que, por ação ou omissão voluntária, negligência ou imprudência, violar direito ou causar prejuízo.

Esta expressão jurídica da responsabilidade civil classifica-se ainda em dois aspectos: a responsabilização civil subjetiva, que decorre da culpa do agente causador do dano, e a responsabilização civil objetiva, nos casos em que a lei expressamente prescinde do requisito de culpa, bastando a existência de um nexo causal (Gagliano; Pamplona Filho, 2009). Em todo caso, para que seja configurada a responsabilidade civil é preciso primeiramente que haja o dano (Gagliano; Pamplona Filho, 2009), ensejador da obrigação em repará-lo.

Deste modo, para o seguro de responsabilidade civil, a ocorrência de dano à terceiro, combinada ou não com a reclamação deste, de acordo com as condições da apólice, configura o sinistro indenizável pela seguradora (SUSEP, 2021).

De maneira geral, segundo a SUSEP (2022), os sinistros correspondem aos eventos previstos e cobertos nas apólices, certificados, bilhetes, contratos ou planos. Já a sinistralidade, de acordo com Pandolfi e Gonçalves (2024), é um indicador prudencial representado pela relação entre o montante de sinistros e o volume de prêmios ganhos.

No âmbito atuarial, o acompanhamento da sinistralidade é considerado fundamental para a sustentabilidade técnica da carteira. A análise de sinistros está associada com uma das

principais atividades de uma companhia de seguros, sob a ótica de Roa e Gonsalves (2015), no que diz respeito à tarefa de administrar os fluxos de prêmios e indenizações entre segurados e seguradores.

Os mesmos autores defendem ainda que o meio econômico é fator importante para a precificação de prêmios, por serem estes influenciados pelas condições de mercado, normas regulatórias, atividade de competidores e nível de renda da população. Sendo assim, é possível que em um determinado meio econômico de comercialização de seguros, a relação entre sinistros e prêmios, isto é, a sinistralidade, possa ser estudada e explicada à luz de fatores pertencentes a este mesmo mercado.

1.1 PROBLEMA DE PESQUISA

A partir dessas considerações introdutórias, indaga-se para fins desta pesquisa: Quais fatores observáveis do mercado segurador brasileiro podem explicar a dinâmica da sinistralidade no ramo do seguro de responsabilidade civil geral?

1.2 OBJETIVOS

1.2.1 Objetivo geral

O objetivo geral do trabalho, consiste, portanto, em analisar, sob uma perspectiva atuarial, a sinistralidade dos seguros de Responsabilidade Civil Geral no Brasil, a partir dos dados regulatórios do mercado segurador nacional, identificando tendências, sazonalidades, heterogeneidades e fatores determinantes para a sinistralidade.

1.2.2 Objetivos específicos

A fim de alcançar o objetivo geral, os objetivos específicos da pesquisa dividem-se em:

- Coletar informações regulatórias de prêmios, sinistros, provisões técnicas e indicadores de capital, no período de novembro/2016 a maio/2025, para o Seguro de Responsabilidade Civil Geral (Ramo 0351);
- Estruturar base de dados consolidada para o ramo estudado, abrangendo todo o mercado supervisionado, em recorte mensal por entidade seguradora;
- Aplicar métodos de análise temporal da sinistralidade no ramo estudado, a fim de investigar a existência de estrutura temporal própria;
- Construir variáveis indicadores de comportamento empresarial, a fim de segmentar as entidades em agrupamentos de mercado;

- Aplicar métodos estatísticos-atuariais para descrever e interpretar o comportamento da sinistralidade no ramo, considerando as heterogeneidades dos agrupamentos do mercado supervisionado.

1.3 JUSTIFICATIVA

A escolha do ramo de seguros de responsabilidade civil justifica-se em razão de seu crescimento em importância no país, tendo acumulado em 2025, até o mês de maio, cerca de 1,7 bilhão de reais em prêmios e 530 milhões de reais em indenizações; um aumento de 8,87% em relação ao mesmo período no ano anterior, de acordo com o CNSEG (2025). O segmento de Danos e Responsabilidades, que inclui os seguros de responsabilidade civil, pagou 27 bilhões de reais em indenizações neste mesmo período; um crescimento de 6,9% segundo os mesmos dados do CNSEG (2025).

Notadamente em relação ao seguro de Responsabilidade Civil Geral, Oliveira e Tavares (2014) argumentam, em linhas gerais, que seu crescimento esteja relacionado com o crescimento econômico do país, isto é, o aumento do produto interno bruto. Em outra linha, Prado (2021) aponta para a potencialidade do seguro de responsabilidade civil em uma "sociedade litigiosa", indicando que exposições em esferas trabalhista e de consumo, bem como novas tecnologias e questões ligadas à diversidade e governança, dentre outros fatores, podem impactar os seguros de responsabilidade civil, alterando assim a demanda por esse instrumento.

Adicionalmente, cabe destacar que o seguro de responsabilidade civil exerce função sobre a eficiência do sistema econômico. Ao transferir riscos de indenizações potencialmente elevadas para o mercado segurador, esse instrumento reduz custos de transação associados à litigiosidade, permitindo que agentes econômicos mantenham sua capacidade produtiva. Neste sentido, Coase (1960) defende que os custos de transação são significativos ao ponto de desincentivar a existência de certas relações comerciais.

Ao mesmo tempo, a existência de responsabilidades, em sentido estrito, asseguradas, incentiva os potenciais causadores de danos a serem mais cuidadosos. Segundo Baker (2013), estes agentes suportam os custos de suas precauções, mas também colhem os benefícios, na forma de uma expectativa de responsabilização menor junto aos possíveis terceiros afetados. A previsibilidade e a liquidez trazidas pelo seguro de responsabilidade podem reduzir os custos da litigância e ampliar a segurança dos agentes econômicos.

Dessa forma, a escolha do ramo 0351 para esta pesquisa justifica-se não apenas por seu crescimento e pela importância no mercado segurador brasileiro, mas também pelo papel que desempenha como mecanismo de eficiência econômica e de estabilidade financeira. A análise

da sinistralidade, sob perspectiva atuarial e regulatória, revela-se importante para compreender os efeitos desse segmento na resiliência do sistema econômico.

2 REFERENCIAL TEÓRICO

Este capítulo fornece a fundamentação teórica que sustenta a pesquisa, abrangendo conceitos normativos e técnicos vinculados ao mercado segurador brasileiro, com destaque para o ramo 0351, referente aos seguros de responsabilidade civil geral. São apresentados, ainda, os fundamentos da análise de séries temporais e dos modelos de vetores autorregressivos, enquanto recursos a serem empregados na compreensão da dinâmica da sinistralidade.

Na sequência, são exploradas as bases conceituais da análise de agrupamento de dados, como instrumento para a caracterização do comportamento das seguradoras, e, por fim, do modelo linear generalizado, aplicado à explicação do mercado em estudo na medida de suas heterogeneidades. A articulação desses referenciais visa à construção de um substrato de consistência teórica capaz de amparar as análises empíricas a serem realizadas.

2.1 CONCEITUAÇÃO TÉCNICO REGULATÓRIA

A presente seção consolida os conceitos técnico-regulatórios indispensáveis à análise da sinistralidade no ramo 0351 (Responsabilidade Civil Geral), a fim de padronização terminológica e aderência às normas vigentes no mercado supervisionado. No ordenamento nacional, a regulação dos seguros cabe ao Conselho Nacional de Seguros Privados (CNSP) e a supervisão e execução normativa à Superintendência de Seguros Privados (SUSEP), cujas resoluções, circulares e manuais disciplinam a elaboração das demonstrações, a constituição de provisões técnicas, a mensuração de capital e o reporte de informações estatísticas e prudenciais pelas seguradoras (CNSP, 2015; SUSEP, 2020; SUSEP, 2021).

Os dados empíricos deste estudo derivam do Sistema de Estatísticas da SUSEP (SES), alimentado por remessas padronizadas do Formulário de Informações Periódicas (FIP/SUSEP), sob regime de competência e com taxonomia oficial de ramos, eventos e contas técnicas. Essa infraestrutura regulatória define, entre outros, os conceitos de prêmio direto, prêmio ganho, sinistro ocorrido e as principais provisões técnicas, bem como os indicadores prudenciais de patrimônio líquido ajustado (PLA) e capital mínimo requerido (CMR), com vistas à comparabilidade intertemporal e entre entidades supervisionadas.

2.1.1 O ramo 0351 – Responsabilidade Civil Geral

A Superintendência de Seguros Privados (SUSEP), autarquia vinculada ao Ministério da Fazenda, é responsável pela regulação e fiscalização do mercado segurador no Brasil,

disciplinando as operações de seguros, resseguros, capitalização e previdência complementar aberta (SUSEP, 2022).

No âmbito da codificação dos ramos de seguro, a SUSEP estabelece que o ramo 0351 corresponde ao Seguro de Responsabilidade Civil Geral (RCG). Esse produto destina-se a garantir ao segurado a cobertura de indenizações devidas a terceiros em razão de danos materiais, corporais ou morais decorrentes de fatos não intencionais, relacionados ao exercício de suas atividades (CNSP, 2021; SUSEP, 2012).

O RCG é classificado entre os seguros de responsabilidade civil facultativos, diferenciando-se de outros ramos, como o de responsabilidade civil de veículos automotores (RCF-V) ou o de responsabilidade profissional, por possuir abrangência mais ampla e adaptável a diferentes contextos de atividade econômica (Oliveira, 2014). O risco coberto é essencialmente de terceirização do prejuízo, em que a seguradora se obriga a ressarcir o segurado das quantias que venha a ser condenado a pagar judicial ou extrajudicialmente (SUSEP, 2021).

2.1.2 Provisões técnicas e conceitos de mensuração no mercado segurador brasileiro

As provisões técnicas representam passivos obrigatórios que as seguradoras devem constituir, com o objetivo de garantir a solvência das obrigações decorrentes dos contratos de seguro. Segundo a Resolução CNSP nº 432/2021, as principais provisões relacionadas aos seguros de danos e responsabilidades são:

- Provisão de Prêmios Não Ganhos (PPNG): corresponde à parcela dos prêmios emitidos que ainda não foi apropriada como receita, refletindo a obrigação da seguradora de prestar cobertura durante o período de risco ainda vigente. Na prática, a PPNG assegura que a companhia possua recursos suficientes para honrar eventos futuros ainda cobertos, mesmo que o prêmio correspondente já tenha sido recebido.
- Provisão de Sinistros a Liquidar (PSL): corresponde aos sinistros já avisados e ainda não pagos, incluindo estimativas de valores que a seguradora deverá desembolsar no futuro. É formada a partir da comunicação de ocorrência do sinistro e deve refletir o valor esperado da indenização bruta de resseguro, acrescido das despesas relacionadas à regulação e liquidação.
- Provisão de Sinistros Ocorridos mas Não Avisados (IBNR): corresponde a sinistros que já ocorreram, mas ainda não foram notificados à seguradora. Essa provisão é fundamental em ramos danos e responsabilidades, em razão do intervalo

potencialmente longo entre a ocorrência do evento danoso e o conhecimento pelo segurado ou seguradora.

De acordo com Sandström (2011), a adequada constituição das provisões técnicas é um componente da solvência. Insuficiências na estimação das reservas podem comprometer a solvência da seguradora, ao passo que excessos podem impactar a competitividade. As provisões técnicas atuais têm relação com o volume de sinistros passados, exceto para a PPNG, de modo que quanto maior a sinistralidade maior possam ser as reservas.

Além das provisões, alguns conceitos regulatórios definidos pela SUSEP e pelo CNSP são centrais para a análise atuarial. Identificam-se primeiramente as definições envolvendo prêmios e sinistros, que se relacionam na mensuração da sinistralidade.

Os prêmios podem ser classificados como diretos ou ganhos. O prêmio direto: corresponde ao valor pago pelo segurado à seguradora, antes de qualquer dedução de resseguro, taxas ou comissões, representando a receita bruta de subscrição (SUSEP, 2021). Já o prêmio ganho corresponde à parcela do prêmio direto apropriada como receita do período, proporcional ao tempo decorrido da cobertura do risco.

O sinistro ocorrido: corresponde ao valor dos eventos danosos efetivamente registrados no período de competência, abrangendo sinistros pagos, sinistros a liquidar e movimentações em IBNR. A SUSEP adota a ótica de competência, exigindo que as seguradoras reconheçam no resultado todos os sinistros incorridos, pagos ou não (SUSEP, 2020).

A sinistralidade, portanto, corresponde ao indicador técnico de desempenho, que expressa a relação entre os sinistros ocorridos e os prêmios ganhos no período de competência, refletindo a proporção dos recursos arrecadados em prêmios e o consumo pela reclamação dos eventos cobertos. A sinistralidade é uma métrica de performance atuarial, visto que taxas persistentemente elevadas de obrigações contra as receitas, podem levar à insuficiência de capital da seguradora.

Em relação ao patrimônio e capital de risco das companhias, a Resolução CNSP n° 432/2021 dispõe dos conceitos de Patrimônio Líquido Ajustado (PLA) e Capital Mínimo Requerido (CMR), empregados nesta pesquisa. O PLA corresponde ao patrimônio líquido contábil ajustado por adições e exclusões regulatórias e o CMR corresponde ao montante de capital regulatório necessário para que a seguradora mantenha suas operações.

De acordo com o CNSP (2021), os ajustes realizados no patrimônio líquido, para fins de mensuração do PLA, visam à apuração dos recursos que estão disponíveis para suportar as oscilações e situações adversas decorrentes das atividades das companhias.

Neste mesmo sentido, o capital mínimo a ser mantido pelas companhias que desejam operar no mercado de seguro, baseia-se no ramo de operação e considera diferentes classes de exposição a risco. Segundo o CNSP (2021), o CMR consiste no maior valor entre o capital-base, preconizado para cada ramo supervisionado, e o capital em risco, calculado para cada agrupamento de produtos em que a companhia opera, de acordo com a equação:

$$CR = CR_{geral} - \min\left[CR_{geral} \times PR;\ (PEF + MV) \times \left(1 - \frac{PD}{2}\right)\right]$$

Na equação do capital em risco, CR_{geral} refere-se ao capital de risco de subscrição que está associado às exposições dentro de um determinado agrupamento de produtos e PR refere-se ao menor percentual de reversão de excedentes financeiros observado para os produtos neste agrupamento. Ademais, $PEF + MV$ refere-se às provisões constituídas para os produtos em questão somadas à mais valia dos ativos correspondentes, e PD refere-se à estimativa do percentual de saída dos segurados desse mesmo grupo de produtos nos próximos três meses.

O CMR foi assim definido para levar em consideração os riscos de subscrição, crédito, mercado e operacional das companhias, seguindo parâmetros do modelo de Capital Baseado em Risco (Risk-Based Capital – RBC). Ao mesmo tempo, o PLA contempla ajustes de qualidade (CNSP, 2021), que levam em consideração o CMR, a fim de permitir suficiência de cobertura para as operações.

Juntos, os níveis de patrimônio líquido ajustado e de capital mínimo requerido fornecem sinais observáveis das escolhas de subscrição, gerenciamento de riscos e apetite a risco das seguradoras, tendo em vista a capacidade de absorção de perdas, representada pelo PLA, e a exposição à variabilidade dessas perdas, representada pelo CMR. Esses elementos podem influenciar os critérios de retenção de riscos e aceitação de negócios pelas entidades, afetando, consequentemente, a sinistralidade ao longo do tempo.

Os conceitos apresentados formam a base técnica de mensuração da atividade seguradora no âmbito desta pesquisa. Ao relacionar fatores como sinistralidade, provisões técnicas e capital requerido, é possível avaliar a solvência das carteiras e os efeitos das escolhas estratégicas das seguradoras (Cummins; Sommer, 1995).

2.2 MODELAGEM QUANTITATIVA

Esta seção traz a conceituação teórica sobre as modelagens quantitativas empreendidas no exame do problema de pesquisa.

2.2.1 Análise de séries temporais

O estudo de séries temporais consiste na análise de conjuntos de observações $\{Y_t\}$, indexadas no tempo $t = 1,2,\dots,T$, com o objetivo de compreender suas propriedades estocásticas, identificar padrões estruturais e fornecer ferramentas de previsão (Box et al. 2016; Hamilton, 1994). Séries temporais apresentam dependência sequencial, refletindo a evolução dinâmica de fenômenos econômicos, financeiros e atuariais.

No campo atuarial, séries temporais podem ser empregadas para avaliar a evolução de indicadores técnicos, como a sinistralidade, permitindo identificar comportamentos cíclicos, detectar mudanças estruturais e realizar projeções de curto e longo prazo (Frees et al, 2014).

Uma série temporal pode ser decomposta em quatro componentes fundamentais:

$$Y_t = T_t + S_t + C_t + \varepsilon_t$$

Onde:

- T_t corresponde à tendência de longo prazo, capturando variações persistentes no nível da série;
- S_t corresponde à sazonalidade, padrões cíclicos que se repetem em períodos fixos (ex.: mensal, trimestral, anual);
- C_t denota o ciclo econômico, flutuações irregulares de médio prazo;
- ε_t é o componente irregular, ou ruído branco, definido como um processo $WN(0, \sigma^2)$

Cleveland et al. (1990) propuseram o modelo de decomposição de séries temporais denominado *Seasonal Trend decomposition using Loess* (STL), que segundo os autores permite separar esses elementos, fornecendo diagnósticos úteis para modelagem e previsão.

Grande parte dos modelos de séries temporais requer a hipótese de estacionariedade, ou seja, que as propriedades estocásticas da série não variem no tempo. Formalmente, um processo $\{Y_t\}$ é fracamente estacionário se:

$$E[Y_t] = \mu, \qquad Var(Y_t) = \sigma^2 < \infty, \qquad Cov(Y_t, Y_{t+h}) = \gamma(h), \qquad \forall t, h$$

Quando séries apresentam tendência ou sazonalidade, técnicas como diferenciação $\nabla Y_t = Y_t - Y_{t-1}$ ou transformação logarítmica são aplicadas para alcançar estacionariedade. Teste Dickey-Fuller Aumentado (ADF) pode ser adotado na verificação dessa propriedade (Hamilton, 1994).

Dickey e Fuller (1979) demonstraram que, sob a hipótese nula de uma dada regressão linear apresentar raiz unitária, o valor estimado do coeficiente de seu preditor segue estatística baseada na estimativa da seguinte regressão:

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=0}^{m} \alpha_i \Delta Y_{t-i} + \varepsilon_t$$

O teste ADF examina a estacionariedade através da estatística de $t = \hat{\delta}/S_{\hat{\delta}}$; onde: $\hat{\delta}$ é o estimador MQO de δ e $S_{\hat{\delta}}$ é o erro padrão dessa estimativa.

Os modelos autorregressivos e de médias móveis constituem a base da análise de séries temporais. O modelo $AR(p)$ é definido por:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2)$$

Já o modelo $MA(q)$ é expresso como:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

A combinação dos dois resulta no modelo $ARMA(p, q)$. Quando a série é não estacionária, introduz-se o operador de diferenciação ∇, obtendo-se o modelo $ARIMA(p, d, q)$, conforme Box & Jenkins (1970):

$$\nabla^d Y_t = \phi_1 \nabla^d Y_{t-1} + \cdots + \phi_p \nabla^d Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-1}$$

Para séries com padrões sazonais, utiliza-se o modelo $SARIMA(p, d, q)(P, D, Q)_s$, que incorpora termos sazonais de ordem s

$$\Phi_P(B^s)\phi_p(B)\nabla^d \nabla_s^D Y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t$$

onde B é o operador defasagem, e $\nabla_s^D = (1 - B^s)^D$ é o operador de diferenciação sazonal.

O modelo SARIMA é utilizado na modelagem de séries temporais que apresentam padrões sazonais recorrentes. Sua principal característica reside na capacidade de capturar simultaneamente componentes autorregressivos, de médias móveis e de integração, tanto em nível quanto em frequência sazonal, permitindo representar estruturas temporais complexas observadas em séries econômicas e financeiras (Box et al. 2016; Hamilton, 1994).

O emprego do modelo SARIMA na análise de sinistralidade neste estudo, visa refletir as variações cíclicas associadas a fatores institucionais, operacionais ou macroeconômicos ao longo do tempo, com efeito sobre a série estudada.

2.2.2 Vetores autorregressivos

O método de vetores autorregressivos (VAR), proposto por Sims (1980) e consiste numa modelagem de "equações simultâneas" – que são regressões pelo método dos Mínimos Quadrados Ordinários (MQO), em que cada variável do sistema é explicada por seus valores defasados e pelos valores das outras variáveis presentes no modelo (Gujarati, 2011).

O modelo VAR generaliza a abordagem univariada ao permitir a análise conjunta de múltiplas séries temporais interdependentes. O VAR descreve o vetor de variáveis endógenas $y_t = (y_{1t}, y_{2t}, \dots, y_{kt})^\top$ como função linear de suas próprias defasagens:

$$y_t = A_1 y_{t-1} + A_2 y_{t-1} + \cdots + A_p y_{t-p} + \varepsilon_t$$

Onde A_i são matrizes de coeficientes $k \times k$ e ε_t é um vetor de ruído com matriz de covariância Σ.

O modelo de vetores autorregressivos é utilizado na análise de sistemas dinâmicos multivariados, permitindo capturar as inter-relações entre múltiplas séries temporais sem a imposição de restrições estruturais a priori, e possibilitando a análise de efeitos dinâmicos e de propagação de choques ao longo do tempo (Sims, 1980).

O emprego do modelo VAR na análise de sinistralidade deste estudo visa investigar relações de interdependência entre fatores observáveis de mercado, a fim de compreender como interagem entre si e influenciam a sinistralidade, oferecendo uma estrutura analítica consistente para a identificação de efeitos dinâmicos.

2.2.3 Análise de agrupamento de dados

A análise de agrupamento, ou *cluster analysis*, consiste em um conjunto de métodos estatísticos multivariados cujo objetivo é particionar um conjunto de n observações $\{x_1, x_2, \dots, x_n\}$ com $x_i \in R_p$, em K grupos (ou clusters), de modo que elementos do mesmo grupo apresentem maior similaridade entre si do que em relação aos demais grupos (Everitt et al., 2011).

Essa técnica é de natureza exploratória, podendo ser aplicada em problemas de classificação, segmentação de mercados e estratificação de riscos em seguros (Gan; Valdez; Huang, 2020).

Formalmente, define-se uma função de dissimilaridade $d: R_p \times R_p \to R_+$, tal que:

$$d(x_i, x_j) \geq 0, \qquad d(x_i, x_j) = d(x_j, x_i), \qquad d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$$

O método k-means (MacQueen, 1967) é o mais difundido entre os algoritmos particionais de agrupamento de dados. Seu objetivo é minimizar a soma das distâncias quadráticas intragrupo:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2$$

Onde μ_k é o centróide do cluster C_k. A solução é obtida iterativamente, alternando-se entre (1) atribuição de observações ao cluster mais próximo e (2) atualização dos centroides. Embora eficiente, o método depende da escolha de K e é sensível a outliers.

Em seguros, a análise de clusters permite classificar seguradoras segundo indicadores como sinistralidade, provisões técnicas e capital regulatório, revelando padrões de competitividade ou prudência (Shi; Shi, 2022). Essa segmentação pode ser utilizada como variável latente em modelos de solvência ou precificação, contribuindo para a compreensão da heterogeneidade do mercado.

2.2.4 Modelos lineares generalizados

Os Modelos Lineares Generalizados (GLM), introduzidos por Nelder e Wedderburn (1972), expandem o modelo de regressão linear ao permitir que a variável resposta Ysiga distribuições pertencentes à família exponencial.

As distribuições da família exponencial empregadas neste estudo são: (1) Poisson, (2) Binomial Negativa, (3) Gama, (4) Gaussiana Inversa e (5) Tweedie.

A distribuição de Poisson é caracterizada pela igualdade entre média e variância e pode ser aplicada em modelagem GLM para capturar a relação das variáveis explicativas com uma taxa média, nesse estudo representada pela sinistralidade. Já a distribuição binominal negativa, por sua vez, é uma generalização da distribuição Poisson para a modelagem de situações em que a variância excede a média, sendo que essa sobredispersão pode estar contida na sinistralidade estudada em razão da alta variabilidade entre os riscos não observados diretamente.

Quanto às demais distribuições, Gama e Gaussiana Inversa podem ser empregadas na modelagem de variáveis contínuas e estritamente positivas, compatível com a taxa de sinistralidade. E a família de distribuições Tweedie, desenvolvida por Bent Jorgensen (1987), é caracterizada por uma relação de potência entre a média e variância, do tipo $Var(Y) = \phi\mu^p$, onde p é um índice de potência, que se definido no intervalo $1 < p < 2$ pode apresentar uma estrutura de processo Poisson composto de severidades Gama, permitindo modelar simultaneamente ocorrência e tamanho das perdas agregadas.

No tocante à modelagem linear generalizada, um GLM é definido por três componentes:

- Componente aleatório: $Y_i \sim F(\mu_i, \phi)$, com F pertencente à família exponencial.
- Componente sistemático: $\eta_i = x_i^\top \beta$, onde x_i é o vetor de covariáveis.

- Função de ligação: $g(\mu_i) = \eta_i$, que conecta a média da resposta μ_i ao preditor linear.

Em aplicações securitárias, é possível incorporar um *offset* na modelagem da variável dependente, representando parâmetro de exposição. Nesse caso, define-se por exemplo:

$$log(\mu_i) = x_i^{\top}\beta + log(E_i)$$

Onde E_i é atributo de exposição da esperança condicional visando permitir que o modelo estime de maneira ajustada ao volume de negócios (Dobson; Barnett, 2018).

A estimação dos coeficientes β é feita pelo método de máxima verossimilhança, utilizando algoritmos iterativos como *Iteratively Reweighted Least Squares* (IRLS). A função de verossimilhança é:

$$L(\beta) = \prod_{i=1}^{n} f_Y(y_i; \theta_i, \phi)$$

Com θ_i relacionado a μ_i via função de ligação.

O emprego do modelo GLM neste estudo relaciona-se com a representação da sinistralidade, cuja distribuição pode ser assimétrica e apresentar variância não constante, visto que o modelo linear generalizado permite especificar a função de ligação entre a média da variável de explicada e a combinação linear das variáveis explicativas.

Por fim, as quatro técnicas de modelagem quantitativa empregadas no estudo aderem-se ao objetivo do trabalho. De maneira consolidada, a análise univariada de séries temporais fornece um panorama do comportamento temporal da sinistralidade estudada, que é ampliado pela inclusão dos fatores de mercado que acompanham sua dinâmica, de acordo com a análise multivariada. A partir daí, a modelagem de agrupamentos classifica comportamentos transversais para os fatores observados, permitindo assim a inclusão de categorias na estimação das efeitos estatísticos sobre a variável de interesse.

3 METODOLOGIA

Este capítulo descreve a metodologia adotada para o desenvolvimento da pesquisa, detalhando as etapas necessárias à análise da sinistralidade no seguro de responsabilidade civil geral no Brasil. Inicialmente, são apresentados o tipo de pesquisa e a base de dados utilizada, proveniente do Sistema de Estatísticas da SUSEP, com a delimitação do período de estudo e das variáveis consideradas. Em seguida, descrevem-se os procedimentos de análise exploratória, os testes preliminares e as modelagens quantitativas empregadas, incluindo decomposição de séries temporais, vetores autorregressivos, técnicas de agrupamento e modelo linear generalizado. Por fim, são expostos os métodos de validação aplicados, com vistas à robustez interpretativa dos resultados a serem obtidos.

3.1 TIPO DE PESQUISA

A abordagem metodológica deste trabalho fundamenta-se na pesquisa experimental, conforme caracterizada por Gil (2002), na medida em que busca aplicar métodos estatísticos e econométricos a dados observados para investigar relações causais e padrões estruturais. A pesquisa experimental possibilita a manipulação e análise sistemática de variáveis, com o intuito de observar os efeitos resultantes e, assim, validar ou refutar hipóteses predefinidas. O termo “experimental” no sentido metodológico ampliado engloba a utilização de técnicas quantitativas aplicadas a uma base consolidada de dados.

A escolha por essa abordagem justifica-se pelo problema de pesquisa em questão, que exige não apenas a descrição de fenômenos, mas também a avaliação de relações entre fatores no ramo de seguros de responsabilidade civil. A pesquisa combina a exploração de dados históricos com a aplicação de modelos quantitativos, de agrupamento e regressão, de modo a gerar evidências objetivas sobre a dinâmica do mercado em estudo.

Complementarmente, foi realizada pesquisa bibliográfica e documental a partir do levantamento e análise de contribuições científicas já publicadas, conforme apresentam Marconi e Lakatos (2017), servindo para fundamentar o quadro teórico e conceitual. No presente trabalho, essa etapa envolveu a consulta a livros, artigos acadêmicos, teses, dissertações e publicações de órgãos reguladores.

3.2 BASE DE DADOS DA PESQUISA

A base de dados utilizada neste estudo é proveniente do Sistema de Estatísticas da Superintendência de Seguros Privados (SES/SUSEP), que consolida mensalmente informações

contábeis e atuariais enviadas de forma compulsória por todas as sociedades supervisionadas. Os dados são disponibilizados em formato *Comma Separated Values* (.csv), o que permite acessibilidade e padronização na coleta de informações.

O SES é uma base de dados abrangente do mercado segurador brasileiro, cobrindo dimensões diversas da atividade, tais como: prêmios, sinistros, provisões técnicas detalhadas, patrimônio e capital baseado em risco, limites de retenção e cessões de resseguro, além de bases específicas para operações de renda e capitalização, dentre outras. No conjunto, as tabelas permitem análises por entidade, grupo econômico, ramo de seguros e unidade federativa. As estatísticas do SES são abastecidas por meio do Formulário de Informações Periódicas (FIP).

3.2.1 Formulário de Informações Periódicas (FIP)

O Formulário de Informações Periódicas é o instrumento regulamentar por meio do qual as instituições supervisionadas transmitem dados periódicos à SUSEP. A exigência está originalmente estabelecida na Circular SUSEP nº 189/2002, que define os prazos e a obrigatoriedade de envio do FIP às empresas autorizadas para operar no mercado segurador brasileiro

O FIP passa por diversas atualizações todos os anos, com revisões dos quadros estatísticos exigidos rastreadas no Manual de Preenchimento FIP/SUSEP, que detalha o que deve ser reportado e em que formato. O FIP é composto por múltiplos quadros estatísticos, que as seguradoras devem compulsoriamente preencher. Esses quadros constituem o Sistema de Estatísticas da SUSEP (SES), que consiste na base adotada nesta pesquisa.

O envio do FIP é realizado por meio de aplicativo oficial “FipSusep.exe”, um sistema que roda em ambiente Windows e armazena os dados em arquivo Access (*.MDB) denominado FIP.MDB. Periodicamente, a SUSEP atualiza o executável e os parâmetros de validação, incorporando novos quadros ou revisando regras de preenchimento A submissão dos dados envolve:

- Preenchimento e validação local: o sistema realiza validações lógicas e alertas (críticas) que devem ser justificadas pela seguradora
- Transmissão via internet: os dados são enviados ao sistema da SUSEP utilizando portas de comunicação específicas
- Protocolo de recebimento: o sistema gera um protocolo que confirma o recebimento dos dados.

A obrigatoriedade de preenchimento e envio do FIP por todas as sociedades seguradoras legais garante que os dados do SES sejam completos, padronizados e homologados. O

detalhamento dos quadros, a granularidade (mensal, por empresa e por ramo) e a cobertura institucional fazem dos dados obtidos pelo FIP/SUSEP uma fonte com confiabilidade estatística e regulatória.

A evolução dos quadros e das validações pela SUSEP reflete mudanças regulatórias, e aprimoramentos contínuos na governança de dados, reforçando também a necessidade de os interpretar corretamente quando da coleta.

3.2.2 Janela de pesquisa, coleta e preparação da base de dados

Serão selecionados os seguintes arquivos do SES, com base no objeto da pesquisa:

- Ses_seguros.csv: dados de prêmios e sinistros por ramo, entidade e período;
- Ses_provramos.csv: dados de provisões técnicas por ramo, entidade e período;
- Ses_seg_prov_det.csv: dados de provisões técnicas por entidade e período;
- Ses_prov.csv: dados de provisões totais por entidade e período;
- Ses_pl_margem.csv: dados de patrimônio e solvência por entidade e período.

Essas bases cobrem o universo de companhias supervisionadas, e serão filtradas de forma a selecionar as observações do ramo 0351 – Responsabilidade Civil Geral, por mês e por entidade no período novembro de 2016 a maio de 2025.

A escolha da janela temporal justifica-se pela combinação de dois fatores principais: (1) atualidade dos dados, a fim de que o estudo reflita as condições mais recentes do mercado; (2) aprimoramento das informações de solvência do mercado, com a inclusão do Capital Mínimo Requerido (CMR) nas estatísticas da base adotada, seguindo as definições do capital de risco dispostas na Circular nº 517/2025 da SUSEP.

Segundo Gil (2017), a delimitação temporal de uma pesquisa deve considerar fatores institucionais que conferem relevância e validade ao recorte adotado. Na mesma linha, Marconi e Lakatos (2010) destacam que a confiabilidade da análise depende da consistência das fontes de dados disponíveis, especialmente quando estas decorrem de alterações normativas. Em complemento, Creswell (2014) observa que eventos regulatórios que alteram a forma de mensuração constituem critérios metodológicos legítimos para a definição da abrangência temporal da pesquisa.

A coleta dos dados será realizada diretamente no portal do SES/SUSEP, seguida de tratamento em linguagem Python para filtragem, padronização e integração das tabelas. A preparação deverá incluir:

1) Filtro do ramo 0351 no arquivo Ses_seguros.csv, e seleção das variáveis: premio_de_seguros, premio_ganho, despesa_resseguros, receita_resseguro, desp_com, sinistro_ocorrido, ppng, psl, ibnr.

2) Definição de chave de referência composta pelo código da entidade supervisionada e pelo período, na tabela filtrada do arquivo Ses_seguros.csv.

3) Filtro a partir da chave de referência nos demais arquivos selecionados, Ses_provramos.csv, Ses_seg_prov_det.csv, Ses_prov.csv e Ses_pl_margem.csv, e seleção das variáveis: NovoPla, CMR, prem_n_ganhos, sinistros_liquidar, sinistros_ibnr.

4) Limpeza de colunas vazias ou redundantes.

5) Exclusão das companhias que não apresentam observações para toda a janela.

6) Agregação das provisões técnicas por companhia.

7) Inclusão do prêmio ganho agregado em todos os ramos por companhia.

8) Consolidação das tabelas filtradas e variáveis incluídas em um único data-frame.

Essa sequência de ações visa permitir a construção de uma base única, estruturada em formato painel, a fim de suportar as análises definidas no objetivo e fundamentadas no referencial.

3.3 DEFINIÇÃO DE VARIÁVEIS

A etapa de definição das variáveis constitui ponto central da pesquisa, pois é por meio dela que se estabelecem os indicadores capazes de traduzir conceitos econômico-atuariais em medidas quantitativas verificáveis. Como ressaltam Marconi e Lakatos (2017), a clareza na definição conceitual e operacional das variáveis é condição para assegurar consistência metodológica e viabilizar comparações adequadas ao longo do tempo e entre diferentes unidades de análise. Neste trabalho, todas as variáveis derivam das estatísticas oficiais divulgadas pela Superintendência de Seguros Privados (SUSEP), filtradas para o ramo 0351 – Seguro de Responsabilidade Civil Geral.

A variável dependente central do estudo é a sinistralidade corrente, medida pela razão entre os sinistros ocorridos e os prêmios ganhos (Ohlsson; Johansson, 2010). Este indicador costuma ser visto em conjunto com a taxa de despesas, medida pela razão entre despesas comerciais e o prêmio emitido, e, juntas, representa a performance técnica do ramo na entidade e no período, permitindo identificar a relação direta entre os eventos cobertos e a capacidade de cobertura proporcionada pelas receitas de prêmios.

No conjunto das variáveis explicativas que estão apresentadas no Quadro 2, encontram-se aquelas que traduzem aspectos da estrutura de solvência e comportamento econômico das seguradoras. A primeira delas é a participação do ramo nas provisões, obtida pela razão entre as provisões técnicas constituídas no ramo 0351 e o total das provisões técnicas da companhia. Este indicador reflete o peso relativo do ramo (Sandström, 2011) na composição do passivo técnico, funcionando como parâmetro da relevância estratégica do ramo na estrutura da empresa.

A segunda variável, denominada alavancagem de subscrição, é calculada pela razão entre os prêmios ganhos no ramo 0351 e o patrimônio líquido ajustado da seguradora. Representa o grau de exposição técnica frente ao capital disponível. Conforme Daykin et al. (1994), a medida relaciona exposição de subscrição com base de capital.

A terceira variável refere-se à cobertura de capital ou folga de solvência, definida como a razão entre o patrimônio líquido ajustado e o capital mínimo requerido (Cummins, Phillips, 2009). Esse indicador mensura a distância regulatória em relação ao limite prudencial estabelecido pela SUSEP, funcionando como métrica de suficiência da cobertura. Quanto maior a folga, maior a capacidade da seguradora em suportar choques adversos sem comprometer sua continuidade operacional.

A variável de interesse seguinte é a retenção de risco, definida pela razão entre prêmios retidos e prêmios brutos emitidos no ramo 0351. Esse indicador mostra a dependência da seguradora em relação ao resseguro. Segundo Vaughan e Vaughan (2014), níveis mais baixos de retenção indicam maior transferência de risco a terceiros.

Outra variável é o crescimento do prêmio, mensurado pela taxa de variação dos prêmios emitidos no ramo 0351 em relação ao período imediatamente anterior. Constitui parâmetro de agressividade comercial, refletindo estratégias de expansão da companhia no mercado. Segundo Cummins e Weiss (2014), o crescimento excessivo pode representar má gestão de riscos e constitui uma das principais causas de insolvência.

Por fim, considera-se a concentração no ramo, definida pela proporção entre os prêmios do ramo 0351 e o total de prêmios emitidos pela seguradora. Trata-se de um indicador de dependência estratégica da empresa em relação ao ramo de responsabilidade civil, podendo revelar perfis diferenciados de exposição de acordo com a especialização ou diversificação das entidades (Cummins; Rubio-Misas, 2006).

O Quadro 2 contém o enunciado de todas a variáveis do estudo. Os índices nas formulações 0351, i e t correspondem, respectivamente, a ramo de seguro, entidade e período.

Quadro 1. Indicadores econômico-atuariais do mercado

Métrica	Fonte bibliográfica	Variável derivada	Enunciado
Desempenho	Ohlsson e Johansson (2010)	Sinistralidade	$\frac{SO_{0351,i,t}}{PG_{0351,i,t}}$
Rentabilidade	Ohlsson e Johansson (2010)	Índice de despesas	$\frac{DESP_{0351,i,t}}{PS_{0351,i,t}}$
Concentração de riscos	Sandström (2011)	Participação do ramo nas provisões	$\frac{PPNG_{0351,i,t} + PSL_{0351,i,t} + IBNR_{0351,i,t}}{PT_{i,t}}$
Exposição a riscos	Daykin et al. (1997)	Alavancagem de subscrição	$\frac{PG_{0351,i,t}}{PLA_{i,t}}$
Suficiência de cobertura	Cummins e Phillips (2009)	Folga de solvência	$\frac{PLA_{i,t}}{CMR_{i,t}}$
Transferência de riscos	Vaughan e Vaughan (2014)	Retenção de risco	$\frac{PS_{0351,i,t} - D.RES_{0351,i,t} + R.RES_{0351,i,t}}{PS_{0351,i,t}}$
Práticas de gestão	Cummins e Weiss (2014)	Crescimento do prêmio	$\frac{PS_{0351,i,t} - PS_{0351,i,t-1}}{PS_{0351,i,t-1}}$
Especialização em riscos	Cummins e Rubio-Misas, (2006)	Concentração no ramo	$\frac{PG_{0351,i,t}}{PG_{i,t}}$

Fonte: Elaboração própria.

Em conjunto, essas variáveis compõem um painel de indicadores que articula dimensões técnicas, regulatórias e estratégicas do mercado analisado. Sua utilização visa permitir investigar, sob a ótica atuarial e quantitativa, a evolução da sinistralidade do ramo 0351, bem

como os fatores internos que podem explicar padrões e tendências, atendendo ao objetivo geral desta pesquisa.

3.4 ANÁLISE EXPLORATÓRIA E TESTES PRELIMINARES

A aplicação da modelagem quantitativa será precedida pela realização de uma análise exploratória dos dados (*Exploratory Data Analysis* – EDA). A EDA constitui um conjunto de procedimentos destinado a compreender a estrutura dos dados, identificar padrões, detectar anomalias e avaliar pressupostos de modelagem. Nesse sentido, trata-se não apenas uma preparação técnica, mas também um instrumento de verificação da qualidade e da consistência da base de dados.

A análise exploratória será conduzida inicialmente por meio de estatísticas descritivas (média, quartis, desvio-padrão, e medidas de assimetria e curtose). Esses indicadores permitem caracterizar a distribuição das variáveis e verificar a existência de discrepâncias ou distorções relevantes. Além disso, a análise gráfica de séries temporais, como ferramentas visuais de apoio, na identificação de tendências, sazonalidades e outliers.

Após as análises iniciais, serão procedidos os testes para subsidiar as modelagens quantitativas, sendo eles: (1) endogeneidade, (2) estacionariedade, (3) normalidade e (4) correlação e multicolinearidade.

3.4.1 Endogeneidade

A presença de endogeneidade em equações de regressão implica correlação entre pelo menos uma variável explicativa e o termo de erro, o que invalida a consistência dos estimadores por MQO. Em dados de painel, uma fonte recorrente de endogeneidade é a simultaneidade entre as variáveis.

A abordagem no tocante à endogeneidade neste trabalho será mediante defasagens dos regressores na forma de variáveis explicativas instrumentais e aplicação do teste de Durbin-Wu-Hausman, que visa avaliar se um determinado regressor x está correlacionado com o erro e da regressão estrutural, de forma que:

$$E[x|e] = 0$$

O teste adota o procedimento de estimação de mínimos quadrados ordinários em dois estágios. No primeiro estágio, a variável explicativa potencialmente endógena é modelada a partir de um instrumento e das demais variáveis explicativas. No segundo estágio, o resíduo do primeiro estágio é incluído no modelo da equação estrutural para a variável dependente, de tal

modo que caso possua coeficiente significativo encontra-se diagnosticada a endogeneidade do regressor correspondente.

Kelejian e Robinson (1998) defendem que defasagens podem ser consideradas instrumentos para variáveis endógenas. A força de tais instrumentos é medida pelo teste F de significância dos coeficientes para a regressão do primeiro estágio. A estatística de F superior a 10 indica instrumento fortes para substituir as variáveis endógenas na regressão estrutural da variável dependente.

3.4.2 Estacionariedade

Em análise de séries temporais, um processo estocástico é dito estacionário quando suas propriedades estatísticas permanecem constantes ao longo do tempo. Nesse sentido, se uma série temporal é não estacionária, só é possível estudar seu comportamento no instante considerado, de modo que o propósito de previsão pode ser de pouco valor prático.

O diagnóstico de estacionariedade das variáveis presentes nesse trabalho incluirá a análise gráfica das funções de correlação (FAC) e correlograma e o teste da raiz unitária de Dickey-Fuller. Conforme Gujarati (2011), a FAC de uma amostra apresenta-se da seguinte forma:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$

Em que $\hat{\gamma}_k$ e $\hat{\gamma}_0$ são, respectivamente a covariância da amostra e a variância da amostra, dadas por:

$$\hat{\gamma}_k = \frac{\sum(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{n-1}$$

$$\hat{\gamma}_0 = \frac{\sum(Y_t - \bar{Y})^2}{n-1}$$

O gráfico de $\hat{\rho}_k$ contra o número de defasagens k é conhecido como correlograma. A escolha de k é uma questão empírica, todavia Gujarati (2011) sugere que possa ser computada a função de correlação de um terço a um quarto da extensão da série temporal. Assim, a apresentação desejada para o correlograma consiste em valores da função e correlação oscilando em torno de zero ao longo de várias defasagens.

Seguindo a mesma linha, o teste da raiz unitária permite investigar a estacionariedade de uma série temporal. Partindo da equação $Y_t = \rho Y_{t-1} + u_t$, com $-1 \leq \rho \leq 1$ e u_t um termo de erro de ruído branco, quando $\rho = 1$, ou seja, a equação acima tiver uma "raiz unitária", trata-se de um processo estocástico não estacionário, conforme Gujarati (2011).

Com efeito, o teste da raiz unitária consiste em verificar a hipótese de $\rho = 1$, no contexto da equação a seguir, em que $\delta = (\rho - 1)$ e Δ é o primeiro operador da diferença. Desse modo, o teste da raiz unitária torna-se o teste da hipótese de $\delta = 0$.

$$\Delta Y_t = \delta Y_{t-1} + u_t$$

Conforme demonstrado por Dickey e Fuller (1979), sob a hipótese nula, o valor estimado t do coeficiente de Y_{t-1} segue a estatística τ . O teste de raiz unitária empregado nesse trabalho é conhecido por teste Dickey-Fuller Aumentado (DFA), que consiste em estimar a seguinte regressão:

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=0}^{m} \alpha_i \Delta Y_{t-i} + \varepsilon_t$$

Cuja estatística de teste equivale a: $t = \hat{\delta}/S_{\hat{\delta}}$; onde: $\hat{\delta}$ é o estimador MQO de δ e $S_{\hat{\delta}}$ é o erro padrão dessa estimativa. O Quadro 3 deve ser usado para a realização do teste DFA; o segmento adotado depende se o modelo for assumido sem constante nem tendência (τ), com constante, mas sem tendência (τ_μ) ou com constante e tendência (τ_τ).

Quadro 2. Estatística do teste DFA para os níveis de significância 0,01, 0,05 e 0,10

Tamanho da amostra	Nível de significância		
	0,01	0,05	0,10
	τ		
25	-2,66	-1,95	-1,60
50	-2,62	-1,95	-1,61
100	-2,60	-1,95	-1,61
250	-2,58	-1,95	-1,62
300	-2,58	-1,95	-1,62
∞	-2,58	-1,95	-1,62
	τ_μ		
25	-3,75	-3,00	-2,62
50	-3,58	-2,93	-2,60
100	-3,51	-2,89	-2,58
250	-3,46	-2,88	-2,57
300	-3,44	-2,87	-2,57
∞	-3,43	-2,86	-2,57
	τ_τ		
25	-4,38	-3,60	-3,24
50	-4,15	-3,50	-3,18
100	-4,04	-3,45	-3,15
250	-3,99	-3,43	-3,13
300	-3,98	-3,42	-3,13

Tamanho da amostra	Nível de significância		
	0,01	0,05	0,10
∞	-3,96	-3,41	-3,12

Fonte: Elaboração própria.

Em caso de ausência de estacionariedade para a série da variável estudada, serão adotados procedimentos de diferenciação ou transformação logarítmica.

3.4.3 Normalidade

A normalidade das séries será examinada mediante análise gráfica de qq-plot e histograma, já mencionados, em conjunto com o teste de Jarque-Bera. A estatística do teste é dada pela equação a seguir e consiste em verificar se a curva de densidade da variável estudada apresenta comportamento simétrico e mesocurtico, compatível com uma curva normal.

$$JB = \frac{n}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right]$$

Onde: S é medida de simetria da amostra, K é medida de curtose da amostra e n é o número de observações.

Sob a hipótese nula de normalidade, JB é distribuído como uma distribuição qui-quadrado com 2 graus de liberdade, de sorte que se o p-valor da estatística de teste acima for menor que o nível de significância, tem-se que os valores em análise não estão distribuídos como uma normal.

3.4.4 Correlação e multicolinearidade

O coeficiente de correlação de Pearson é empregado para medir a intensidade e a direção de associações lineares entre as variáveis contínuas. Essa análise permite identificar potenciais redundâncias entre variáveis e oferece indícios preliminares sobre interdependências a serem investigadas na modelagem multivariada.

A presença de multicolinearidade é examinada mediante o fator de incremento da variância (FIV), conforme proposto por Belsley et al. (1980), dado pela equação a seguir.

$$FIV = \frac{1}{1 - r_{ij}^2}$$

Com base no coeficiente de determinação no denominador acima r_{ij}^2, quando a colinearidade aumenta no sentido da unidade tem-se que o FIV aumenta também, tendendo ao infinito. Valores de FIV maiores que 10 são usualmente interpretados como indicativos de

multicolinearidade, ensejando reavaliação do conjunto de variáveis ou aplicação de técnicas de regularização.

3.5 ANÁLISE DE ESTIMAÇÃO DA MODELAGEM QUANTITATIVA

Nesta seção estão apresentados os modelos e métodos quantitativos empregados para fins de alcance dos objetivos desta pesquisa.

3.5.1 Análise univariada de séries temporais

Nesta etapa estará contemplada a aplicação de modelagem univariada de séries temporais à variável dependente central do estudo – a sinistralidade corrente do ramo 0351. O objetivo dessa abordagem é identificar padrões estruturais na série histórica, tais como tendência e sazonalidade, além de fornecer previsões de curto prazo que sirvam de base para comparações com modelos multivariados posteriores.

Conforme Box e Jenkins (2016), a modelagem de séries temporais deve seguir três etapas principais: (1) identificação do modelo apropriado, com base na análise das funções de autocorrelação (ACF) e autocorrelação parcial (PACF); (2) estimação dos parâmetros e (3) verificação da adequação do modelo por meio da análise de resíduos. Para as etapas 1 e 2 deste roteiro será empregado o procedimento automatizado *auto-arima*, implementado na biblioteca *pmdarima* do Python.

No presente estudo, a modelagem univariada empregada será o modelo SARIMA. O teste de Ljung-Box será aplicado aos resíduos para verificar a ausência de autocorrelação serial e o teste de Goldfeld-Quandt também será aplicado aos resíduos para verificar presença de heterocedasticidade.

O emprego de modelagem univariada justifica-se pelo papel de tais técnicas no diagnóstico inicial do comportamento temporal das séries. Como analisa Hamilton (1994) mesmo em análises que posteriormente evoluem para modelos multivariados, a compreensão da dinâmica univariada pode separar variações estruturais de movimentos conjunturais e avaliar a presença de memória longa ou de padrões sazonais persistentes.

Assim, ao final desta etapa de modelagem, visa-se obter uma primeira caracterização estatística da sinistralidade do ramo 0351, identificando tendências e sazonalidades relevantes e fornecendo subsídios para a etapa seguinte da pesquisa, em que serão considerados modelos multivariados e explicativos.

3.5.2 Análise multivariada de séries temporais

Enquanto a análise univariada visa permitir compreender os padrões internos das séries temporais, a dinâmica do mercado faz necessário considerar inter-relações entre múltiplas variáveis que evoluem conjuntamente no tempo. Sendo assim, nesta etapa será procedida modelagem multivariada de séries temporais, mais precisamente o modelo de vetores autorregressivos (VAR).

O modelo VAR, proposto por Sims (1980), constitui uma generalização multivariada do modelo autorregressivo univariado (AR). A atratividade do VAR reside no fato de que todas as variáveis são tratadas como endógenas, permitindo investigar relações dinâmicas sem impor, a priori, restrições unidirecionais. No contexto do objetivo da pesquisa, é possível que variáveis como alavancagem de subscrição, retenção de risco e folga de capital possam tanto influenciar como ser influenciadas pela sinistralidade.

Para estimação da modelagem VAR serão seguidas as etapas: (1) determinação da ordem de defasagem mediante critérios de informação AIC e BIC, (2) estimação de parâmetros via mínimos quadrados ordinários, (3) análise dos resíduos.

O uso da modelagem VAR neste estudo justifica-se por sua capacidade de capturar interdependências entre múltiplos fatores internos das seguradoras, respeitando a natureza dinâmica dessas relações no estudo de choques, ciclos e interações entre as variáveis. Assim, a modelagem multivariada por VAR complementa a análise univariada, permitindo apontar fatores de mercando determinantes na compreensão sistêmica da sinistralidade.

3.5.3 Análise de agrupamento (*clustering*)

A partir das análises univariada e multivariada de séries temporais, essa etapa da pesquisa incorpora uma etapa de análise de agrupamentos (clustering) com o propósito de identificar perfis das seguradoras que atuam no ramo 0351. A análise de cluster constitui uma técnica estatística exploratória que busca particionar um conjunto de observações em grupos de acordo com medidas de similaridade, de forma que empresas pertencentes a um mesmo cluster apresentem características mais próximas entre si do que em relação às demais.

Essa etapa visa permitir capturar heterogeneidades estruturais entre as entidades, relacionadas, como, por exemplo, diferenças no apetite ao risco, na alocação de provisões, no grau de retenção de prêmios. Ao incorporar essas diferenças na modelagem da etapa seguinte, busca-se evitar a imposição de uma estrutura homogênea sobre o mercado, reconhecendo que distintas estratégias empresariais podem afetar a evolução da sinistralidade.

Para a aplicação da análise de cluster, as variáveis selecionadas como base de agrupamento serão aquelas definidas no Quadro 2. Esses fatores refletem tanto a posição técnica

quanto a orientação estratégica das seguradoras e, portanto, oferecem um conjunto adequado de dimensões para a segmentação.

O processo de análise seguirá duas etapas principais:

- Padronização das variáveis: adoção da normalização *range-bound* para fins de comparabilidade em escalas distintas;
- Definição de agrupamento e número ótimo de *cluster*: adoção do algoritimo *k-means clustering*, que minimiza a soma das distâncias quadráticas dentro de cada *cluster*, resultando em agrupamentos baseados na proximidade em espaço multidimensional.

Os agrupamentos definidos serão interpretados sob a ótica do mercado de seguros, buscando caracterizar grupos de seguradoras como, por exemplo, de perfil conservador (baixa alavancagem e alta folga de solvência) ou agressivo (alta alavancagem, baixa retenção de resseguro e crescimento acelerado de prêmios).

Essa caracterização visa a contribuir para o entendimento do setor e fornecer uma variável categórica a ser abordada na modelagem quantitativa da etapa seguinte, com vistas a ampliação da capacidade explicativa do(s) modelo(s) para a sinistralidade

Dessa maneira, a análise de cluster representa um elo metodológico intermediário entre a análise de séries temporais e a modelagem de regressão, na etapa seguinte, com o propósito de assegurar que as diferenças estruturais entre empresas não sejam negligenciadas, abordando, de forma realista, a heterogeneidade do mercado segurador no ramo em estudo.

3.5.4 Análise de modelo linear generalizado

A etapa final de modelagem quantitativa deste estudo consiste na aplicação de Modelos Lineares Generalizados (*Generalized Linear Models* – GLM), cuja finalidade é explicar a sinistralidade do ramo 0351 em função das variáveis internas das seguradoras e de acordo com os clusters identificados previamente. A escolha pelo GLM justifica-se pela flexibilidade em lidar com variáveis que podem não seguir distribuição normal, além de permitir a introdução de funções de ligação adequadas e termo de *offset*.

No presente trabalho, o *offset* será definido pelo prêmio ganho, de modo que a sinistralidade seja modelada em termos relativos, ajustando o efeito da exposição ao risco.

A modelagem nessa etapa assume a forma:

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \log(PG_{0351,i})$$

Em que, as variáveis explicativas $(x_{1i}, \dots, x_{ki})$ incluem as variáveis definidas no Quadro 2. Quanto à escolha da família de distribuição, cinco especificações serão exploradas: (1) Poisson, (2) Binomial negativa e (3) Gama, (4) Gaussiana Inversa e (5) Tweedie. A seleção entre as distribuições será guiada por critérios de ajuste (AIC, BIC) e pela análise dos resíduos.

Dessa forma, nesta etapa final, a modelagem quantitativa permitirá avaliar de maneira integrada o impacto das variáveis internas nas seguradoras sobre a sinistralidade.

3.6 AVALIAÇÃO DOS MODELOS ESTIMADOS

A etapa final da metodologia consiste nas técnicas de validação dos modelos estimados. As validações serão conduzidas de acordo com a modelagem quantitativa empregada.

Para as modelagens quantitativas será adotada comparação por critérios de informação AIC e BIC. Para a modelagem de *cluster analysis*, será analisada a validade interna dos agrupamentos pelo índice *silhouette*, proposto por Rousseeuw (1986), que considera simultaneamente a distância média entre os pontos de um *cluster* $(a(i))$ e a menor distância média entre os pontos de outros *clusters* $(b(i))$, enquanto medidas de coesão e separação conforme a equação:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

.

Por fim, os resultados empíricos serão discutidos a partir dos objetivos traçados, conectando evidências estatísticas às questões centrais da pesquisa: tendência estrutural da sinistralidade no ramo 0351, heterogeneidades estruturais no mercado e impactos de fatores observáveis sobre a sinistralidade.

4 RESULTADOS E DISCUSSÕES

Neste capítulo são discutidos os resultados obtidos com a exploração empírica da base de pesquisa, conforme os procedimentos descritos na metodologia e de maneira alinhada aos objetivos geral e específicos traçados no estudo.

4.1 CARACTERIZAÇÃO DA BASE DE DADOS

A base coletada para fins desta pesquisa compreende 4284 observações, em formato tabular e em recorte mensal, referentes aos dados das entidades que operam no ramo 0351 – Seguro de Responsabilidade Civil Geral. As informações são de prêmios, despesas, sinistros, provisões técnicas e patrimônio reportadas pelas companhias à SUSEP.

Figura 1. Primeiras três observações da base de dados

Base de dados: 4284 observações e 14 variáveis

	0	1	2
ID	201612_1414	201612_1571	201612_1627
damesano	2016-12-01 00:00:00	2016-12-01 00:00:00	2016-12-01 00:00:00
coenti	1414	1571	1627
cogrupo	1225	1230	345
premio_de_seguros	1266865.23	2304977.26	529743.4
premio_ganho	1517975.0	4518235.26	603996.8
desp_com	333761.96	640867.75	128950.84
sinistro_ocorrido	-297713.11	-414326.18	24332.98
NovoPla	59053745.08	43257098.12	61453514.72
CMR	40752585.24	23758177.06	17029693.54
provisoes_0351	14102801.95	21862111.16	9542262.53
provisoes_totais	265947463.42	230518747.57	124767930.0
premio_ganho_total	17036156.74	23984037.65	6835269.33
risco_retido	1034305.71	-1131177.83	389292.08

Fonte: Elaboração própria.

De acordo com a Figura 1, a base consolidada após a coleta dos dados de pesquisa contempla as variáveis e categorias discriminadas no Quadro 3.

Quadro 3. Identificação das variáveis coletadas.

Nome da variável	Dado(s) coletado(s) e processado(s)
ID	Chave de referência definida no tópico 2 da seção 3.2.2
damesano	Data em formato mês-ano para cada observação
coenti	Código de entidade definido pela SUSEP
cogrupo	Código de grupo econômico das entidades definido pela SUSEP, não utilizado na pesquisa
premio_de_seguros	Prêmio direto no ramo 0351 por mês e ano e por entidade
premio_ganho	Prêmio ganho no ramo 0351 por mês e ano e por entidade
desp_com	Despesa comercial no ramo 0351 por mês e ano e por entidade

sinistro_ocorrido	Sinistro ocorrido no ramo 0351 por mês e ano e por entidade
NovoPla	PLA por mês e ano e por entidade
CMR	CMR por mês e ano e por entidade
provisoes_0351	Montante de provisões PPNG, PSL e IBNR referentes ao ramo 0351 por mês e ano e por entidade
provisoes_totais	Montante de provisões totais por mês e ano e por entidade
premio_ganho_total	Montante de prêmio ganho total por mês e ano e por entidade
risco_retido	Montante de prêmio direto líquido de despesas e receitas com resseguro por mês e ano e por entidade

Fonte: Elaboração própria.

As entidades seguradoras abrangidas na pesquisa encontram-se consolidadas no Quadro 4. Foram removidas durante às etapas de coleta as observações das empresas que não apresentavam dados para toda a janela estabelecida na metodologia.

Quadro 4. Lista de empresas seguradoras integrantes da base de dados coletada.

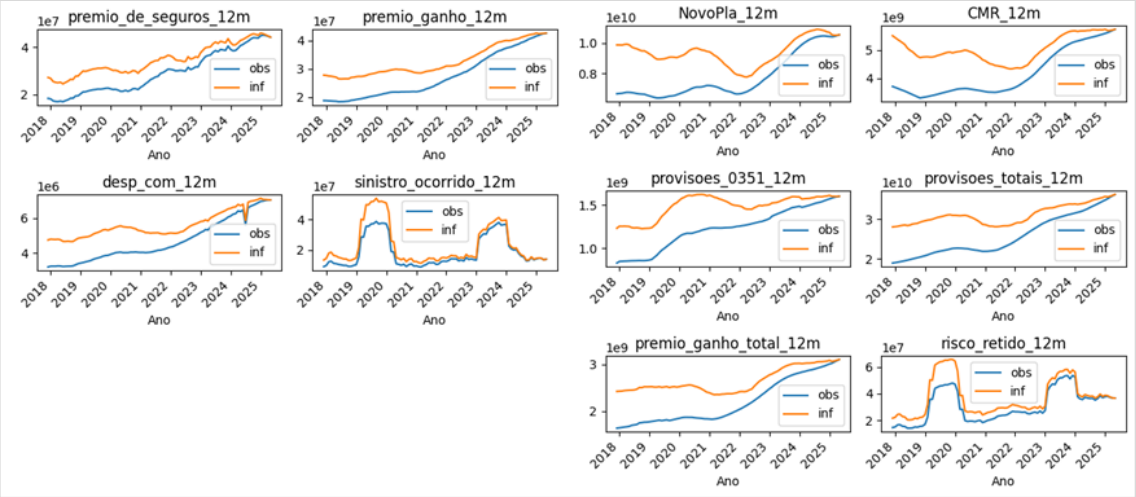
Código	Entidade seguradora
8737	AIG SEGUROS BRASIL S.A.
2798	AKAD SEGUROS S.A.
6467	ALFA SEGURADORA S.A.
6211	ALIANÇA DO BRASIL SEGUROS S.A.
5177	ALLIANZ SEGUROS S.A.
5819	ALLSEG SEGURADORA S.A.
2461	AUSTRAL SEGURADORA S.A.
2852	AXA SEGUROS S.A.
1414	BERKLEY INTERNATIONAL DO BRASIL SEGUROS S/A
5312	BRADESCO AUTO/RE COMPANHIA DE SEGUROS
6785	BRASILSEG COMPANHIA DE SEGUROS
5631	CAIXA SEGURADORA S.A.
6513	CHUBB SEGUROS BRASIL S.A.
5045	COMPANHIA DE SEGUROS ALIANÇA DA BAHIA
5193	COMPANHIA DE SEGUROS PREVIDÊNCIA DO SUL - PREV.
5690	COMPANHIA EXCELSIOR DE SEGUROS
4669	FAIRFAX BRASIL SEGUROS CORPORATIVOS S.A.
3727	FAIRWAY SEGUROS S.A.
6122	FATOR SEGURADORA S/A
5908	GENERALI BRASIL SEGUROS S.A.
6793	GENTE SEGURADORA S.A.
1571	HDI GLOBAL SEGUROS S.A.
6572	HDI SEGUROS S.A.
5843	INDIANA SEGUROS S.A.
5321	ITAU SEGUROS S.A.
6921	KOVR SEGURADORA S.A.
6238	MAPFRE SEGUROS GERAIS S.A.
6602	MITSUI SUMITOMO SEGUROS S.A.
5886	PORTO SEGURO COMPANHIA DE SEGUROS GERAIS
3069	POTTENCIAL SEGURADORA S.A.
1627	SAFRA SEGUROS GERAIS S.A.

Código	Entidade seguradora
2950	SANCOR SEGUROS DO BRASIL S.A.
6751	SEGUROS SURA S.A.
5720	SOMPO SEGUROS S.A.
5991	SWISS RE CORPORATE SOLUTIONS BRASIL SEGUROS S.A.
6190	TOKIO MARINE SEGURADORA S.A.
5118	TRADITIO COMPANHIA DE SEGUROS
1970	UNIMED SEGUROS PATRIMONIAIS S.A.
5185	YELUM SEGUROS S.A.
5941	ZURICH BRASIL COMPANHIA DE SEGUROS
5495	ZURICH MINAS BRASIL SEGUROS S.A.
6564	ZURICH SANTANDER BRASIL SEGUROS S.A.

Fonte: Elaboração própria.

De acordo com a SUSEP, as informações da base são de natureza contábil. Sendo assim, para fins de comparabilidade da análise intemporal, bem como suavização de potenciais oscilações nos valores brutos reportados pelas companhias, as séries monetárias foram agregadas em 12 meses e inflacionadas à mesma base, em maio/2025. A análise gráfica de séries temporais acumuladas e inflacionadas, em valores agregados médios, encontra-se na Figura 2.

Figura 2. Séries temporais agregadas observadas e inflacionadas entre nov/2016 e mai/2025



Fonte: Elaboração própria.

Do ponto de vista descritivo, observa-se que as séries temporais compostas apresentam comportamento variável ao longo do período analisado, com episódios de elevação acentuada, para as series acumuladas de sinistro ocorrido e risco retido, sendo este último calculado a partir do prêmio emitido, líquido de receitas e despesas com resseguros. A ocorrência de flutuações relevantes na série agregada de sinistros reportados condiz com a natureza estocástica das

perdas securitárias, em razão de fatores econômicos, institucionais e comportamentais não observados diretamente.

Observa-se também período de redução suave em despesas comerciais, prêmios ganhos totais e provisões totais das companhias, que se opõe ao aumento, contemporâneo, nas provisões reportadas para o ramo em estudo, indicando um momento de especialização do mercado; o que condiz com a emissão da Circular SUSEP nº 517, de 27 de julho de 2021, enquanto principal instrumento que dispõe acerca dos produtos do ramo de responsabilidades no mercado de seguros brasileiro.

A partir de tais séries apresentadas foram derivadas a variável dependente: sinistralidade média no ramo 0351, e as variáveis explicativas a serem utilizadas neste estudo, conforme definidas na metodologia.

4.2 PRÉ-PROCESSAMENTO

Antes da realização das modelagens estatísticas e econométricas, procedeu-se a uma etapa de pré-processamento da base de dados, com o objetivo de verificar a consistência das taxas derivadas utilizadas no estudo. Isso porque a construção de indicadores derivados a partir de dados contábeis pode produzir valores instáveis quando ocorrem inconsistências de reporte, variações abruptas em denominadores ou mudanças pontuais na estrutura operacional das entidades. Essas situações podem resultar em taxas excessivamente distorcidas, que não refletem o comportamento econômico subjacente das operações de seguro do ramo estudado.

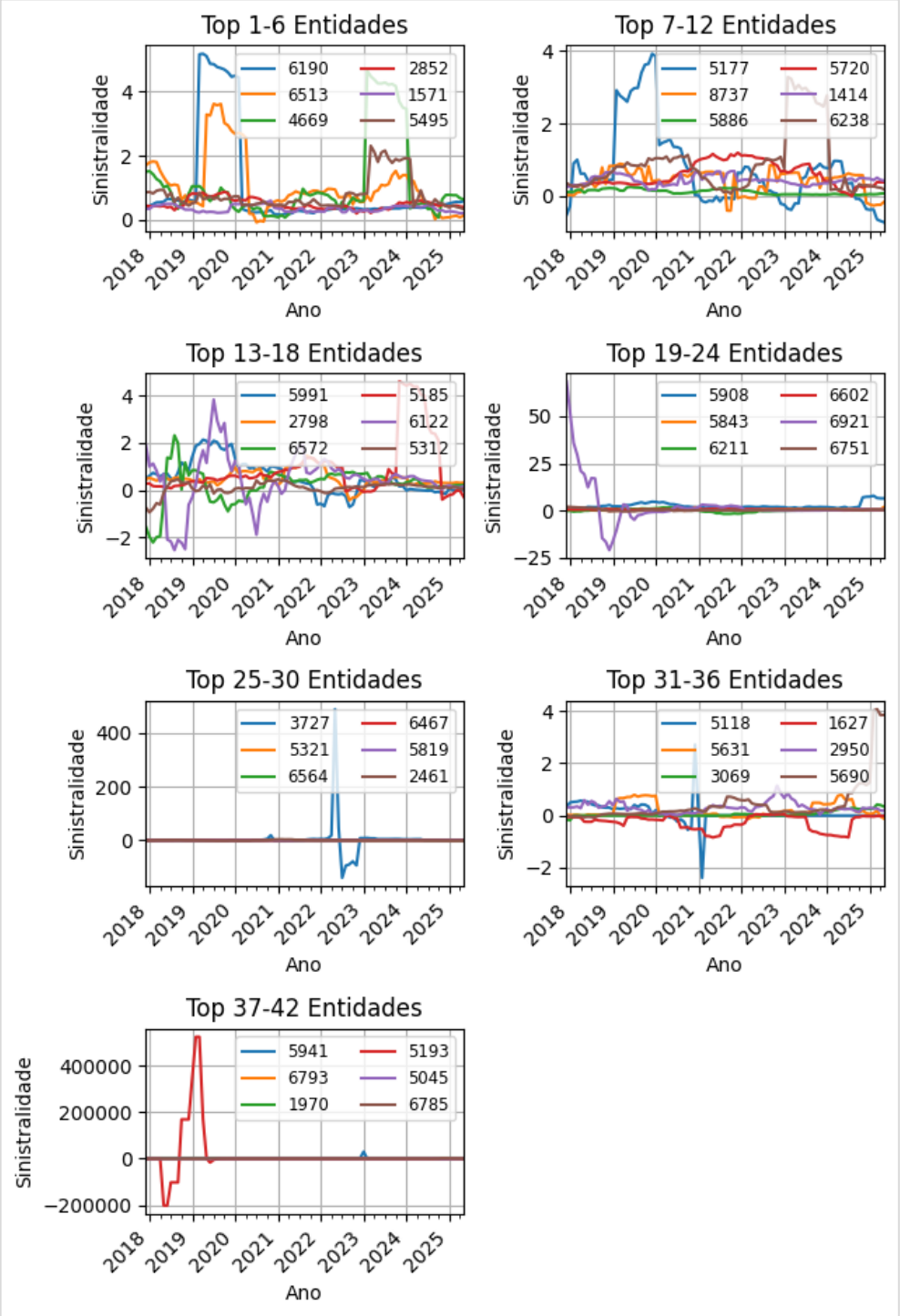
Diante disso, foi conduzida análise exploratória das séries derivadas por meio de inspeção gráfica, procedimento da etapa inicial de análise de dados para identificação de padrões atípicos, inconsistências de medição e possíveis problemas de qualidade da informação (Wilkinson, 2005). A partir dessa análise, verificou-se que determinadas entidades apresentavam taxas que extrapolavam limites economicamente plausíveis, a exemplo de registros para a sinistralidade que atingiam ordens de magnitude extremamente elevadas.

Posteriormente, foram examinados também as distorções para cada variável explicativa vista em dispersão contra a sinistralidade. As análises gráficas estão nas Figuras 3 e 4.

Observou-se que tais comportamentos não se manifestavam de forma pontual, sugerindo inconsistências estruturais nos dados reportados pelas entidades envolvidas ou na própria escala operacional dessas companhias dentro do ramo analisado. Em situações desse tipo, a manutenção de observações pode tão somente distorcer a estrutura estatística da base, especialmente quando tais valores decorrem de problemas de mensuração ou de baixa representatividade econômica (Barnett e Lewis, 1994).

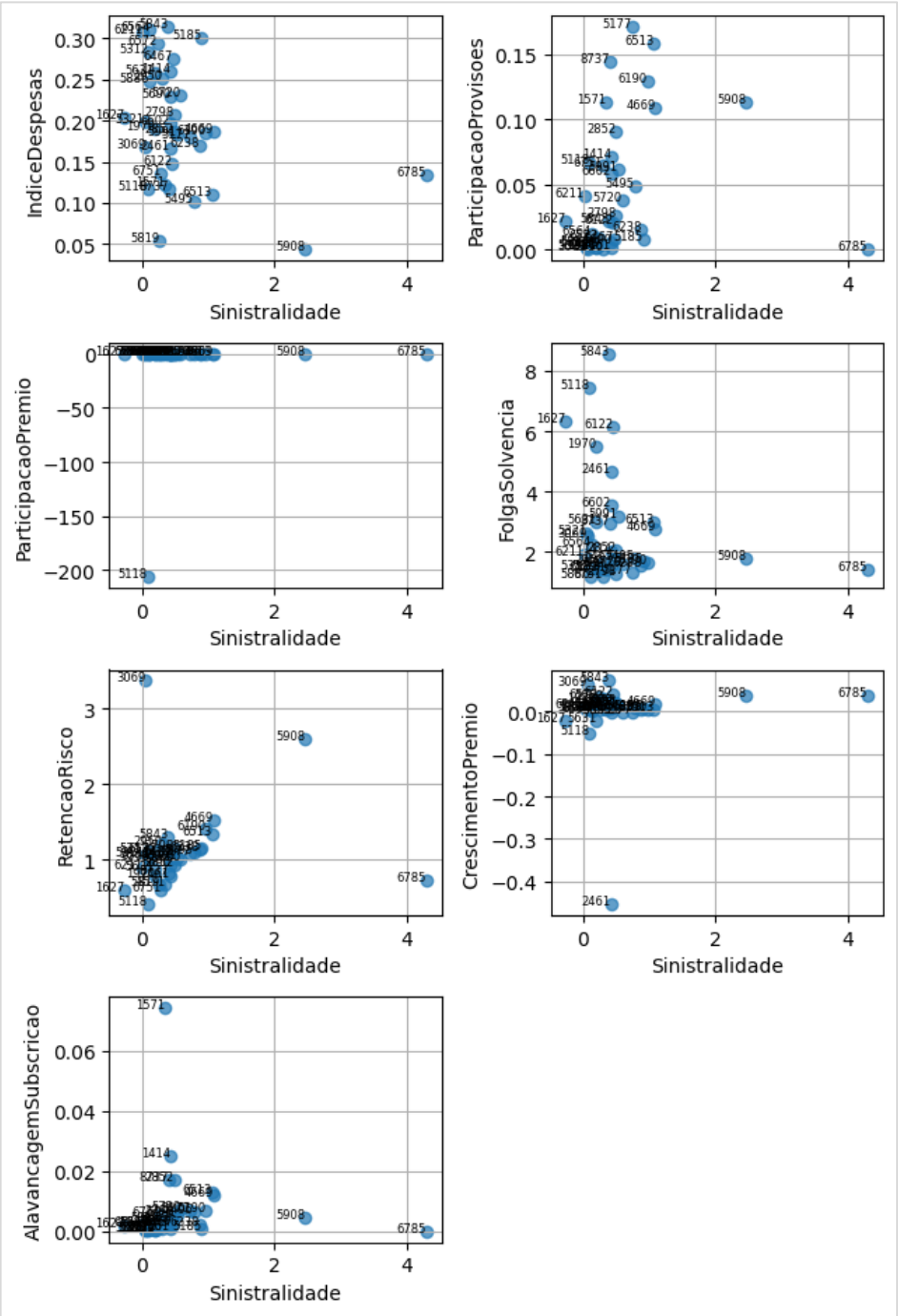
A fim de verificar a representatividade econômica, a referida análise gráfica de comportamento aberrante foi precedida do ordenamento de todas as entidades com base no volume de prêmios emitidos no ramo analisado, a fim de considerar a influência da estrutura de mercado sobre a estabilidade estatística dos indicadores observados.

Figura 3. Análise gráfica de comportamento aberrante na taxa de sinistralidade.



Fonte: Elaboração própria.

Figura 4. Análise gráfica de comportamento aberrante em taxas explicativas



Fonte: Elaboração própria.

Com base nesses critérios, foram excluídas integralmente da amostra as entidades cujas séries apresentavam padrões sistematicamente inconsistentes nas taxas derivadas. No total, 11 entidades apresentavam séries com padrões sistematicamente inconsistentes, tendo sido, portanto, removidas da base coletada, conforme o Quadro 5.

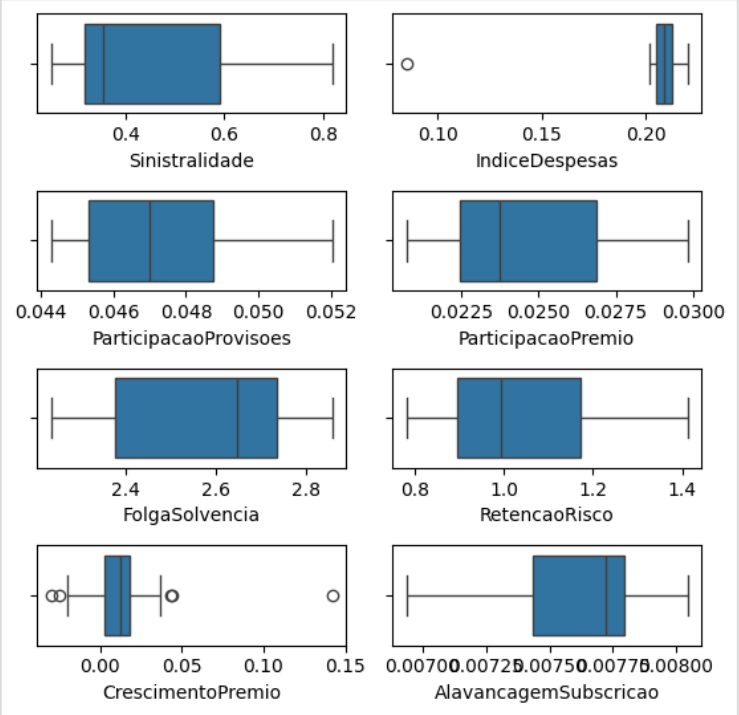
Quadro 5. Entidades removidas da base coletada

Código	Entidade seguradora
2461	AUSTRAL SEGURADORA S.A.
6785	BRASILSEG COMPANHIA DE SEGUROS
5045	COMPANHIA DE SEGUROS ALIANÇA DA BAHIA
5193	COMPANHIA DE SEGUROS PREVIDÊNCIA DO SUL - PREV.
3727	FAIRWAY SEGUROS S.A.
5908	GENERALI BRASIL SEGUROS S.A.
6793	GENTE SEGURADORA S.A.
6921	KOVR SEGURADORA S.A.
3069	POTTENCIAL SEGURADORA S.A.
5118	TRADITIO COMPANHIA DE SEGUROS
5941	ZURICH BRASIL COMPANHIA DE SEGUROS

Fonte: Elaboração própria.

Ainda no âmbito do pré-processamento da base de pesquisa, considerando o universo de dados retidos foi conduzida a verificação de valores extremos, segundo análise gráfica de box-plots disposta na Figura 5. A presença de valor extremo inferior para a variável Índice de Despesas foi tratada por método de winsorização conforme estabelecido na metodologia.

Figura 5. Análise gráfica de valores extremos nas variáveis do estudo



Fonte: Elaboração própria.

4.3 TESTES PRELIMINARES

Com base nos dados consolidados após as etapas de coleta e pré-processamento, foi conduzido um conjunto de testes preliminares com o objetivo de avaliar propriedades estatísticas relevantes das variáveis consideradas na modelagem empírica.

Em particular, de acordo com a metodologia, foram examinados, de maneira sequencial, a endogeneidade das variáveis explicativas, a estacionariedade das séries temporais, a normalidade das distribuições empíricas e à presença de multicolinearidade entre os regressores.

Na avaliação de endogeneidade das variáveis explicativas utilizou-se como instrumento a primeira defasagem de cada variável potencialmente endógena. O uso da primeira defasagem como variável instrumental implica supor que os valores passados podem estar relacionados com os valores contemporâneos dos regressores, sem, contudo, estar relacionado com o erro da equação estrutural.

De acordo com o Tabela 1, as variáveis explicativas Participação em Provisões, Folga de Solvência, Retenção de Riscos e Alavancagem de Subscrição foram substituídas por suas defasagens, em razão do diagnóstico de endogeneidade.

Tabela 1. Teste de endogeneidade das variáveis explicativas

Variável explicativa	Durbin-Wu-Hausman[1]	Força do instrumento[2]
Índice de despesas	0,0584	103,2572
Participação em provisões	0,0000	2895,8621
Participação em prêmios	0,6201	12532,2698
Folga de solvência	0,0078	7958,3212
Retenção de riscos	0,0078	321,2375
Crescimento de prêmios	0,7120	0,3100
Alavancagem de subscrição	0,0004	307,4880

[1] p-valor do teste t de significância do coeficiente do resíduo do primeiro estágio sobre equação do segundo estágio

[2] estatística do teste F para relevância geral do coeficiente da variável instrumental no primeiro estágio

Fonte: Elaboração própria.

A análise de endogeneidade é sucedida pelo exame de estacionariedade das séries temporais, através do teste aumentado de Dickey-Fuller. A variável dependente apresenta estacionariedade sem necessidade de transformação, o mesmo para a variável explicativa Crescimento do Prêmio. As ocorrências de raiz unitária foram tratadas mediante transformações de diferenciação e *detrending*. As taxas transformadas e o diagnóstico de estacionariedade encontra-se no Tabela 2.

Tabela 2. Teste de estacionariedade das séries temporais

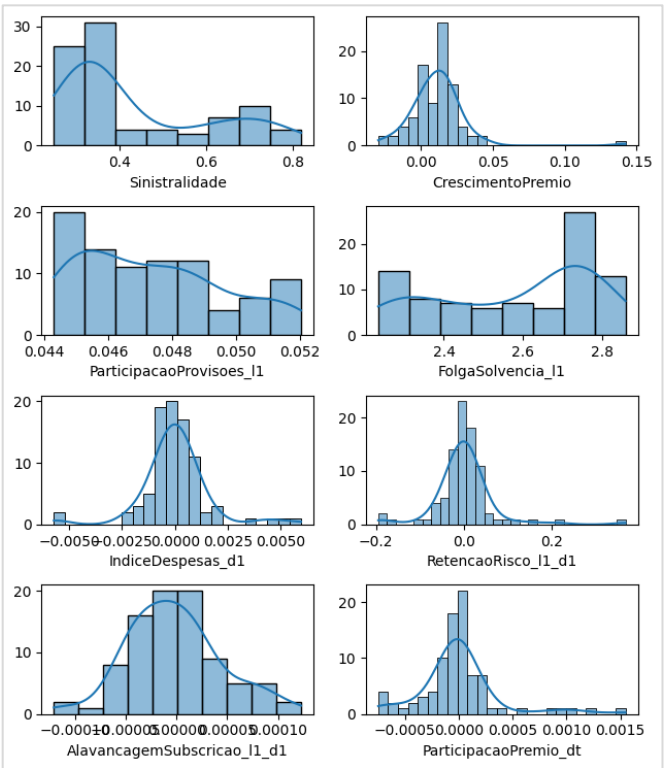
Variável explicativa	Dickey-Fuller[1]	Transformação
Sinistralidade	0,0180	Nenhuma
Crescimento do prêmio	0,0003	Nenhuma
Participação de provisões	0,0002	Defasagem (1)
Folga de solvência	0,0463	Defasagem (1)
Índice de despesas	0,0000	Diferença (1)
Retenção de risco	0,0000	Defasagem (1), Diferença (1)
Alavancagem de subscrição	0,0002	Defasagem (1), Diferença (1)
Participação do prêmio	0,0077	*Detrending*

[1] p-valor do teste ADF

Fonte: Elaboração própria.

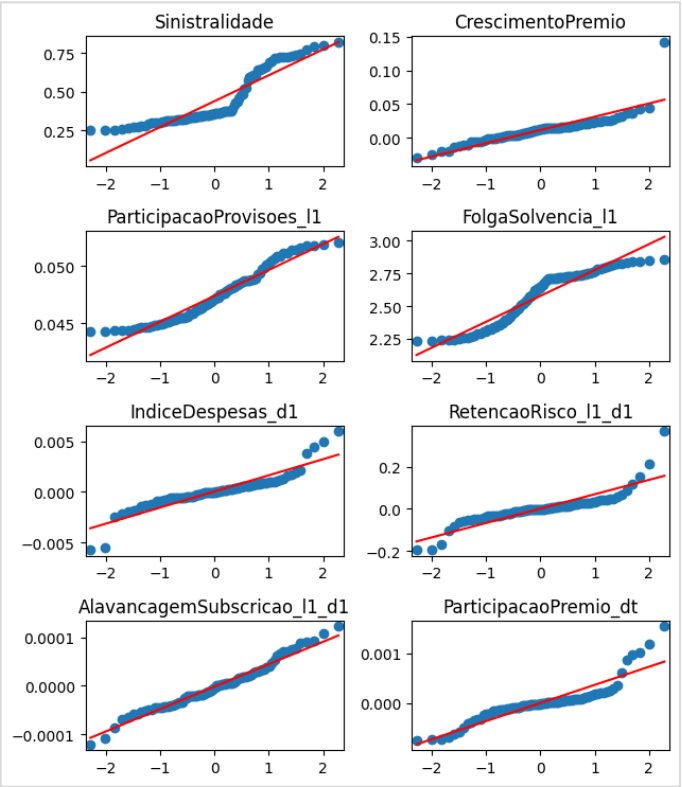
Com o diagnóstico e tratamento de ambas a endogeneidade e a estacionariedade, as séries em estudo são analisadas mediante procedimento gráfico para verificação de normalidade. De acordo com as figuras 6 e 7, três das taxas derivadas apresenta caudas pesadas e comportamento que se distancia da normalidade, em particular a variável dependente, Sinistralidade, e as variáveis explicativas Participação em provisões e Folga de solvência, ambas em primeira defasagem.

Figura 6. Análise gráfica de normalidade das distribuições empíricas – histograma.



Fonte: Elaboração própria.

Figura 7. Análise gráfica de normalidade das distribuições empíricas – qq-plot.



Fonte: Elaboração própria.

Por fim, encerrando os testes preliminares, é analisada a presença de multicolinearidade entre os regressores através do Fator de Incremento da Variância. De acordo com o Tabela 3, as taxas desenvolvidas ocupam níveis de multicolearidade não nocivos para o modelo.

Tabela 3. Análise da presença de multicolearidade

Variável explicativa	FIV
Crescimento do prêmio	1,11
Participação de provisões	1,56
Folga de solvência	1,47
Índice de despesas	1,08
Retenção de risco	1,05
Alavancagem de subscrição	1,17
Participação do prêmio	1,07

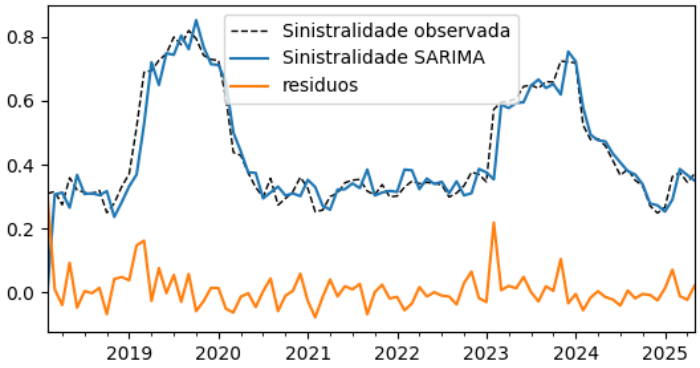
Fonte: Elaboração própria.

As transformações realizadas sobre a base de dados permitem que as modelagens empíricas da sinistralidade preservem as condições estatísticas para análise econômica de seus efeitos.

4.4 A DINÂMICA TEMPORAL DA SINISTRALIDADE MÉDIA NO RAMO (MODELO SARIMA)

O objetivo da modelagem temporal univariada para a Sinistralidade visa investigar se a série apresenta dinâmica temporal própria, que permita ser explicada por fatores observáveis de mercado ou se é puramente aleatória. A modelagem empreendida foi um modelo SARIMA (2,0,1) (2,0,0) [12]. Os parâmetros foram adotados mediante algoritmo de seleção automática baseado em critério de informação AIC e caracterizam um modelo sem diferenciação, com componente não sazonal autorregressivo em 2 lags e erro de previsão dependente da média móvel de ordem1; e sazonalidade autorregressiva também em 2 lags, em ciclo de 12 meses sem dependência de erro dos ciclos anteriores.

Figura 8. Valores observados e ajustados para a sinistralidade média do modelo SARIMA



Fonte: Elaboração própria.

A Tabela 4 permite a comparação entre a sinistralidade média observada no ramo 0351 e a sinistralidade média ajustada pela modelagem SARIMA. O modelo ajustado apresenta três variáveis significativas sendo elas os termos autorregressivos com sazonalidade em defasagem de um e dois períodos e o termo de variância do erro, que representa a parcela não explicada da variância da série temporal em estudo. Já os termos não sazonais do modelo não são significativos, o que implica que os padrões sazonais da sinistralidade são mais dominantes em explicar a dinâmica temporal da série do que os termos de curto prazo não sazonais.

Tabela 4. Resumo da estimação do modelo SARIMA

Termos	Coeficientes	Erro-padrão	P>|z|
AR.L1	0.8169	0.961	0.396
AR.L2	0.1799	0.962	0.852
MA.L1	0.3124	0.956	0.744
AR.S.L12	-0.5623	0.097	0.000

Termos	Coeficientes	Erro-padrão	P>|z|
AR.S.L24	-0.3801	0.108	0.000
Sigma2	0.0022	0.000	0.000

Fonte: Elaboração própria.

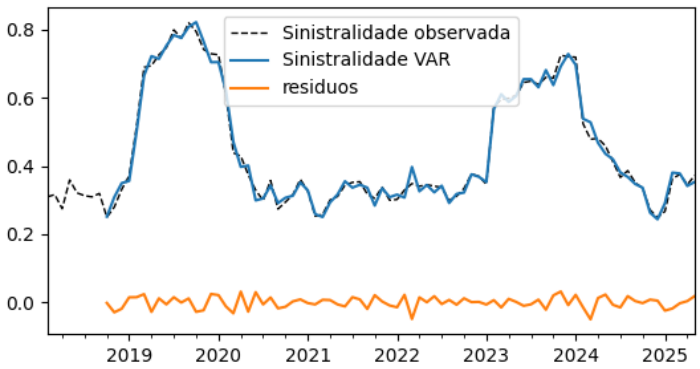
O diagnóstico da modelagem SARIMA apresenta ainda resíduos homocedásticos (teste Goldfeld-Quandt: 0,98; p-valor: 0,96) e não correlacionados (teste Ljung-Box: 0,02, p-valor: 0,89), indicando boa especificação.

A conclusão obtida com a modelagem é de que a sinistralidade média no ramo 0351 apresenta persistência temporal, sugerindo que seu comportamento não é puramente aleatório. O resultado é consistente com a literatura, visto que variáveis econômicas tendem a apresentar autocorrelação e dependência temporal (Hamilton, 1994). Adicionalmente, os mercados de seguro podem apresentar ciclos de subscrição (Cummins, 1987) nos quais indicadores, como a sinistralidade, exibem padrões persistente no tempo. A estrutura temporal da sinistralidade sugere ainda que pode ser influenciada por fatores observáveis no mercado.

4.5 A RELAÇÃO ENTRE A SINISTRALIDADE E FATORES OBSERVÁVEIS DO MERCADO (MODELO VAR)

Ao concluir que a sinistralidade estudada apresenta estrutura de dependência temporal, o objetivo da modelagem multivariada é identificar como fatores observáveis de mercado se relacionam com essa dinâmica. A modelagem empreendida foi um modelo VAR (lag=8). O parâmetro de defasagem foi determinado mediante algoritmo de seleção automática baseado no critério de informação AIC. O modelo regressa cada uma das variáveis derivadas no estudo contra as demais variáveis e suas respectivas defasagens.

Figura 9. Valores observados, ajustados e resíduos do modelo VAR



Fonte: Elaboração própria.

Conforme a Figura 10, a sinistralidade média do mercado obtida pela modelagem VAR ajusta-se próximo da sinistralidade média observada. Dentre as séries temporais estudadas, e suas oito defasagens, cada, regredidas sobre a série temporal da sinistralidade, apresentam coeficientes estatisticamente significativos para o modelo os fatores Alavancagem de subscrição, Participação em Provisões e Índice de Despesas, além de defasagens da própria Sinistralidade, conforme a Tabela 5.

Tabela 5. Resumo da estimação do modelo VAR

Termos significativos	Coeficientes	Erro-padrão	P>|z|
L1.Sinistralidade	0,8466	0,209	0,000
L4.Sinistralidade	0,8532	0,282	0,002
L1.AlavancagemSubscricao_l1_d1	-1012,5263	355,311	0,004
L7.AlavancagemSubscricao_l1_d1	690,4151	312,728	0,027
L2.ParticipacaoProvisoes_l1	393,1099	167,837	0,019
L3.ParticipacaoProvisoes_l1	-419,7189	208,242	0,043
L7.IndiceDespesas_d1	-15,9397	7,437	0,032
L8.IndiceDespesas_d1	-13,0322	6,333	0,039

Fonte: Elaboração própria.

De acordo com a Tabela 5, o choque de 1 ponto percentual sobre o Índice de Despesas em primeira diferença nos períodos 7 e 8, resulta em uma redução de 0,16 e 0,13, respectivamente, sobre o valor atual da sinistralidade média. O achado condiz com o funcionamento esperado do mercado para o seguro de responsabilidade civil, visto que parte das despesas comerciais está relacionada à regulação e defesa do sinistro reclamado, podendo contribuir para a mitigação de perdas, a posteriori.

Também apresentam efeitos negativos sobre a sinistralidade: a taxa de Participação em Provisões, em nível no período 4, e a Alavancagem de subscrição, em diferença no período 2. Para estas, o choque de 0,1 ponto percentual, considerando o domínio das variáveis, resultaria em diminuição de 0,42 e 1,01, respectivamente, no valor da sinistralidade observada. Todavia, os mesmos choques em momentos diferentes sobre essas variáveis, geram efeitos de sinais opostos sobre a sinistralidade atual.

Desta feita os resultados indicam que a exposição a riscos no ramo, em fração do patrimônio, pode ter efeitos temporais dinâmicos sobre o desempenho técnico (Cummins; Sommer, 1996). Efeito negativo, no curto-médio prazo, com a diversificação do capital, e efeito positivo, no médio-longo prazo, com maior assunção de riscos. Ademais, os efeitos dinâmicos são semelhantes com a proporção de reservas técnicas, podendo refletir, em períodos diferentes,

tanto prudência atuarial como também expectativa de perdas, o que pode contribuir para a relação bidirecional do indicador com a sinistralidade.

A conclusão obtida com o modelo VAR reforça o caráter autorregressivo da série temporal da sinistralidade e aponta variáveis que acompanham o comportamento da sinistralidade ao longo do tempo.

4.6 A MODELAGEM ESTATÍSTICA DA SINISTRALIDADE (MODELO GLM)

Ao reconhecer variáveis de mercado que apresentam associação estatística com a estrutura temporal da sinistralidade, o objetivo da modelagem GLM é obter modelo estatístico descritivo para o fenômeno, de modo a indicar os fatores que explicam estatisticamente a sinistralidade. Foram empreendidas cinco modelagens lineares generalizadas, com base nas funções exponenciais: (1) Poisson, (2) Binomial Negativa, (3) Gamma, (4) Gaussiana Inversa e (5) Tweedie. Os modelos especificados atendem à seguinte estrutura:

$$log(\mu_i) = x_i^{\top}\beta + log(P_i)$$

Onde μ_i é a esperança condicional da sinistralidade, cujo comportamento segue distribuição pertencente à família exponencial, $x_i^{\top}$são as séries temporais observáveis no mercado do mercado que acompanham a dinâmica da sinistralidade e P_i é o prêmio médio emitido sem os efeitos da inflação, que expõe a seguradora aos riscos cobertos. O parâmetro de exposição, inserido na forma de *offset* da modelagem, confere pesos proporcionais para a esperança da sinistralidade nos períodos com maiores volumes de transação.

A comparação entre os modelos foi feita mediante critério de seleção AIC, de acordo com Tabela 6, onde também se encontra consolidado o diagnóstico das características dos resíduos, tendo sido empreendidos os testes de heterocedasticidade de Golfeld-Quandt e de Autocorrelação, de Ljung-Box. Os resultados empíricos para os cinco modelos sugerem que não há evidência significativa de heterocedasticidade ou autocorrelação nos resíduos, o que indica boa especificação da dinâmica estrutural da sinistralidade.

Tabela 6. Comparação entre modelagens GLM para a sinistralidade média

Modelo GLM	AIC	Goldfeld-Quandt[1]	Ljung-Box[2]
Poisson	136,183	0,959	0,395
Binomial Negativa	171,792	0,910	0,444
Gamma	-249,637	0,677	0,524

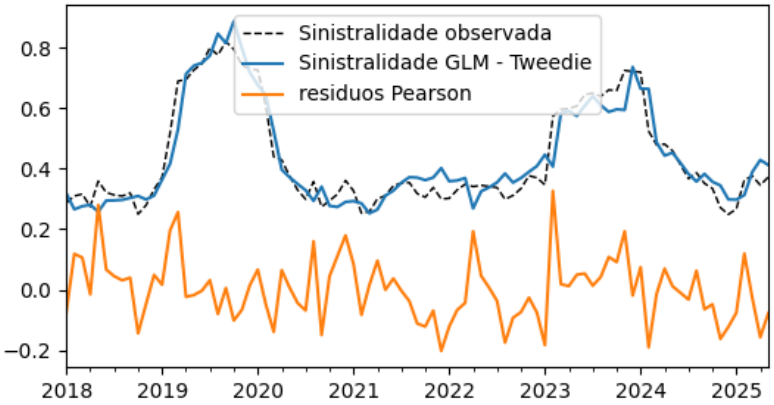
Modelo GLM	AIC	Goldfeld-Quandt[1]	Ljung-Box[2]
Gaussiana Inversa	-240,745	0,357	0,583
Tweedie	-252,049	0,862	0,467

[1] p-valor do teste Goldfeld-Quandt para heterocedasticidade dos resíduos
[2] p-valor do teste Ljung-Box para aucorrelação dos resíduos

Fonte: Elaboração própria.

O Apêndice A fornece as tabelas resumo detalhadas para cada família exponencial modelada. A modelagem mais bem-sucedida, de acordo com o critério de seleção AIC, descreve a sinistralidade média como pertencente à uma distribuição Tweedie. O fator de potência, p, empregado foi 1,5 e dessa forma a distribuição Tweedie se apresenta como um processo composto Poisson-Gamma, o que justifica o forte potencial em capturar simultaneamente frequência e severidade de perdas agregadas atuariais, encapsulados na taxa de sinistralidade média. A Figura 11 apresenta o ajuste do modelo GLM-Tweedie.

Figura 10. Valores observados, ajustados e resíduos Pearson do modelo GLM-Tweedie



Fonte: Elaboração própria.

Os coeficientes estatisticamente significativos para o modelo da sinistralidade média atribuem as variáveis explicativas: Participação no Prêmio, Participação em Provisões e Retenção de Riscos, bem como ainda a defasagem de um período da própria Sinistralidade, conforme indicada pelo modelo VAR, de acordo com o Tabela 7.

Tabela 7. Resumo da estimação GLM-Tweedie

Termos significativos	P>|z|[1]	Coeficiente	Efeito multiplicativo	Variação percentual da sinistralidade
Intercepto	0,000	-15,7219	0,8545	-14,55%
ParticipacaoPremio	0,000	-61,6305	0,5399	-46,01%
Sinistralidade_l1	0,000	2,5852	1,0262	2,62%
ParticipacaoProvisoes_l1	0,001	-39,3960	0,6744	-32,56%
RetencaoRisco_l1	0,000	-0,9697	0,9903	-0,97%

[1] valores 0,000 estão arredondados para fins de apresentação
Fonte: Elaboração própria.

Por se tratar de modelo GLM, os coeficientes estimados mediante função de ligação logarítmica possuem interpretação multiplicativa sobre a média condicional da sinistralidade, de modo que o exponencial, e^{β_k} , do coeficiente, β_k, indica o fator de variação esperado da sinistralidade associado a uma variação unitária na variável explicativa, mantendo constantes os demais fatores. Considerando o domínio das variáveis explicativas, o Tabela 7 adota a variação de 1 ponto percentual no estudo da magnitude do efeito, e, por ser multiplicativo, adota a diferença percentual no estudo do sinal.

Os resultados indicam que o aumento de 0,01 na proporção do prêmio ganho no ramo reduz a sinistralidade média em cerca de 46%. Esse efeito pode ser interpretado como evidência de especialização na subscrição de riscos, não captado na modelagem VAR, por meio da qual as empresas mais especializadas no ramo tentem a assumir riscos mais selecionados e homogêneos, impactando positivamente o desempenho técnico. Observa-se efeito semelhante, em menor magnitude, para a retenção de riscos, também não captada pela modelagem VAR, mas que pode reiterar o efeito de estratégias de subscrição mais criteriosas sobre redução da sinistralidade.

Ainda nessa linha, o aumento de 0,01 na proporção de reservas técnicas no ramo, observadas no período anterior, também diminuem a sinistralidade atual, o que já havia sido identificado na dependência temporal dos fatores.

Ademais, o valor passado da sinistralidade tem efeito positivo sobre a sinistralidade observada de acordo com o modelo. O aumento de 0.01 na taxa sugere elevação de 2,62% para o período seguinte, o que reforça a presença de persistência temporal do indicador.

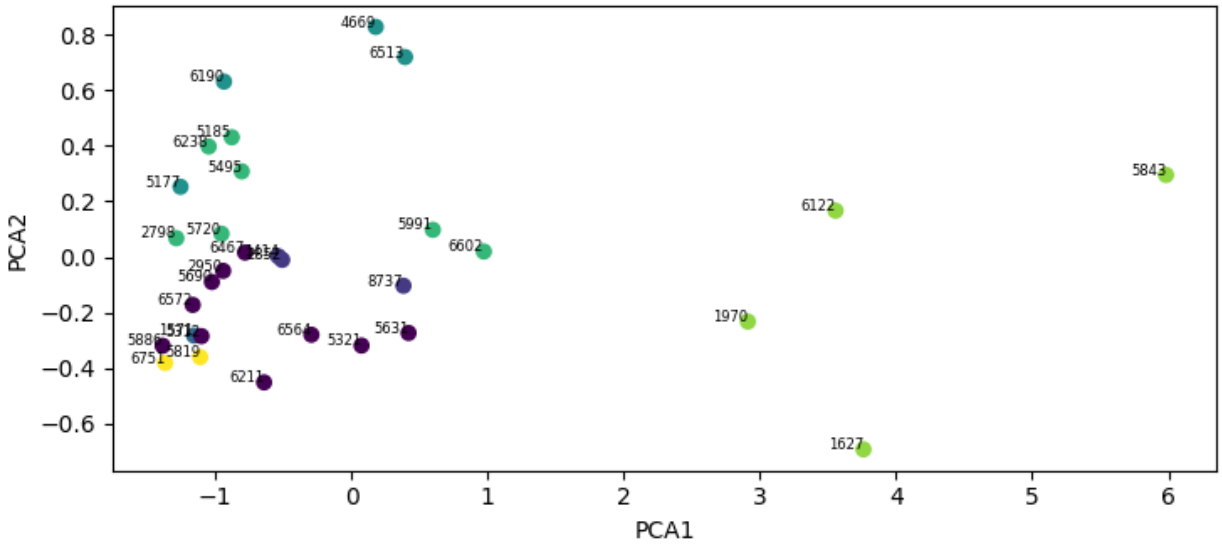
A modelagem linear generalizada da sinistralidade permite concluir que diferentes características de mercado estão associadas estatisticamente a níveis maiores ou menores de sinistralidade.

4.7 A HETEROGENEIDADE NO RAMO 0351 (MODELO CLUSTER)

Com a identificação de variáveis do mercado associadas ao nível de sinistralidade, o objeto da modelagem de agrupamento consiste em examinar se tais características são uniformes a todas as entidades no ramo ou se são consistentes com diferentes estruturas de risco e escalas de operação.

O algoritmo de agrupamento modelado, KMEANS (7), segmentou as 31 entidades seguradores resultantes no estudo em 7 grupos heterogêneos baseados em *proxies* de comportamento empresarial, que consistem nas mesmas taxas médias definidas na metodologia, calculadas ao longo de toda a janela temporal do estudo.

Figura 11. Representação gráfica dos agrupamentos de mercado no ramo 0351

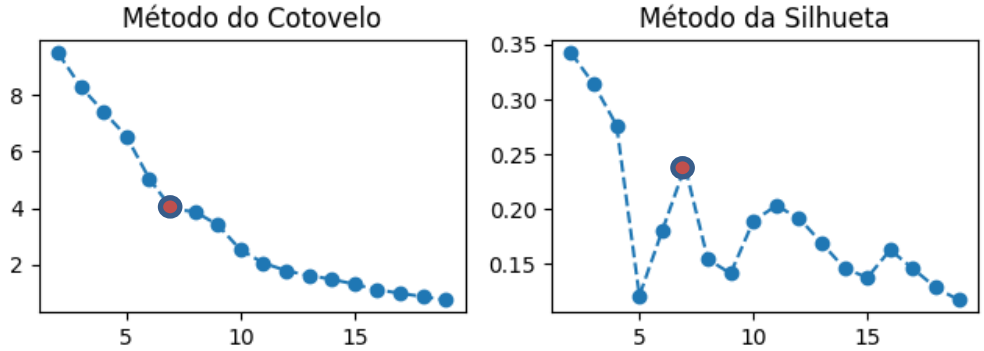


Fonte: Elaboração própria.

A base de *proxies* mencionada foi previamente transformada em escala min-max, para que a distância entre elementos do mesmo grupo reflita dados em mesma dimensão. A representação gráfica da dispersão dos grupos entre as entidades na Figura 11 foi obtida por meio de redução da dimensionalidade através de método PCA e visa indicar os agrupamentos selecionados pelo algoritmo Kmeans.

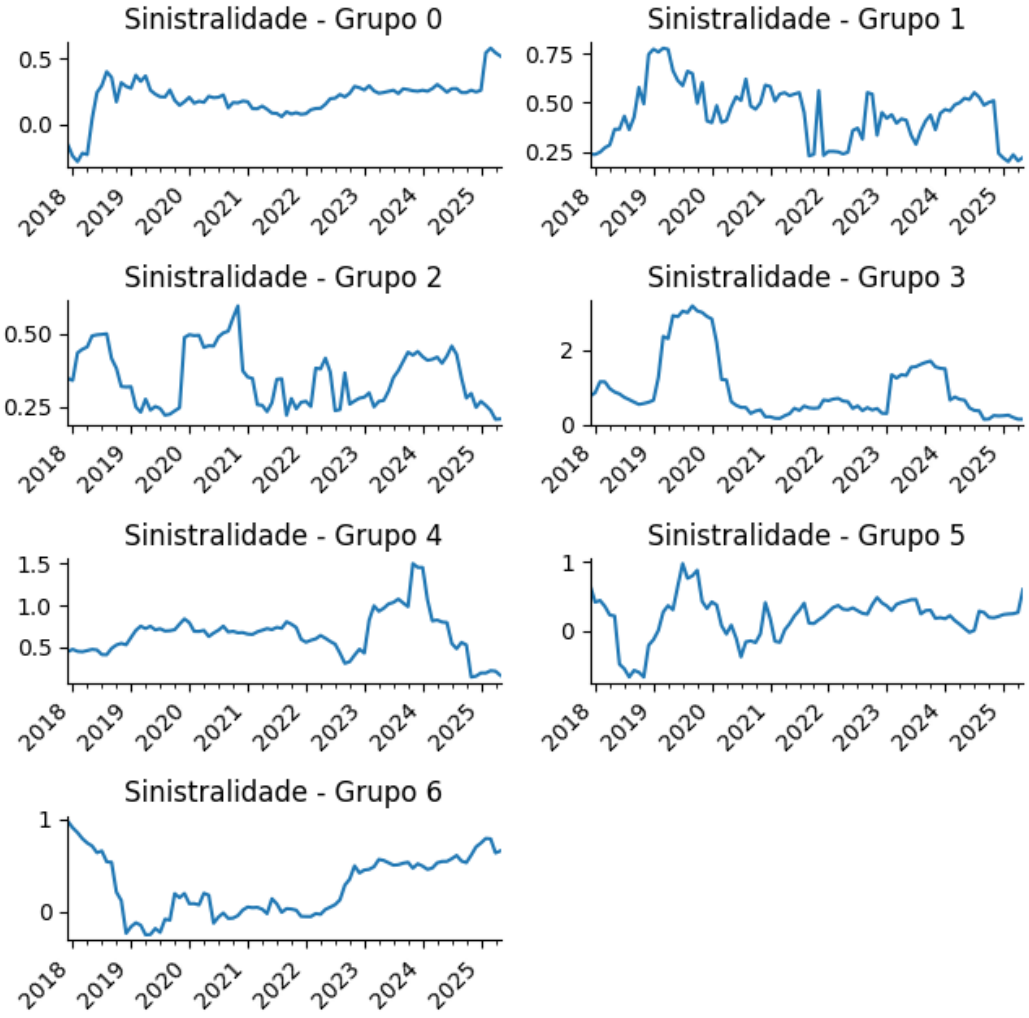
A seleção do número de agrupamentos foi feita a partir de combinação entre o método de cotovelo e o método da silhueta, de acordo com o ponto destacado na Figura 12.

Figura 12. Representação gráfica dos métodos de cotovelo e silhueta para a base agrupada



Fonte: Elaboração própria.

Figura 13. Análise gráfica da sinistralidade média por grupo



Fonte: Elaboração própria.

A partir do mercado segmentado, é possível observar na Figura 13 que cada agrupamento apresenta comportamento típico para a sinistralidade média observada. Destaca-se da análise gráfica a semelhança entre a dinâmica temporal da sinistralidade média no grupo 3 com a sinistralidade média agregada no ramo, examinada nos demais resultados, indicando que este grupo concentra as principais empresas que comercializam o produto, em termos de escala e volume de negócios.

A modelagem de agrupamento permite concluir que o mercado tem estrutura heterogênea, em que diferentes segmentos operam diferentes características para a sinistralidade.

De acordo com a modelagem GLM-Tweedie, já apresentada, é possível examinar os efeitos de tais características que explicam a sinistralidade média. Sendo assim, foram implementados 7 novos modelos GLM-Tweedie, um para cada agrupamento de mercado, com o objetivo de observar as heterogeneidades do ramo. Cada modelo foi precedido pelo diagnóstico e tratamento de endogeneidade entre os regressores, e sucedido pelo diagnóstico de heterocedasticidade e autocorrelação dos resíduos, conforme Tabela 8.

Tabela 8. Comparação e diagnósticos dos resíduos para os modelos de agrupamento

Modelo GLM - Tweedie	Log-Verossimilhança	Goldfeld-Quandt[1]	Ljung-Box[2]
Grupo 0	127,133	0,022	0,023
Grupo 1	87,002	0,375	0,000
Grupo 2	130,101	0,407	0,006
Grupo 3	10,294	0,575	0,004
Grupo 4	67,889	0,069	0,069
Grupo 5	76,010	0,000	0,237
Grupo 6	100,376	0,453	0,000

[1] p-valor do teste Goldfeld-Quandt para heterocedasticidade dos resíduos
[2] p-valor do teste Ljung-Box para aucorrelação dos resíduos

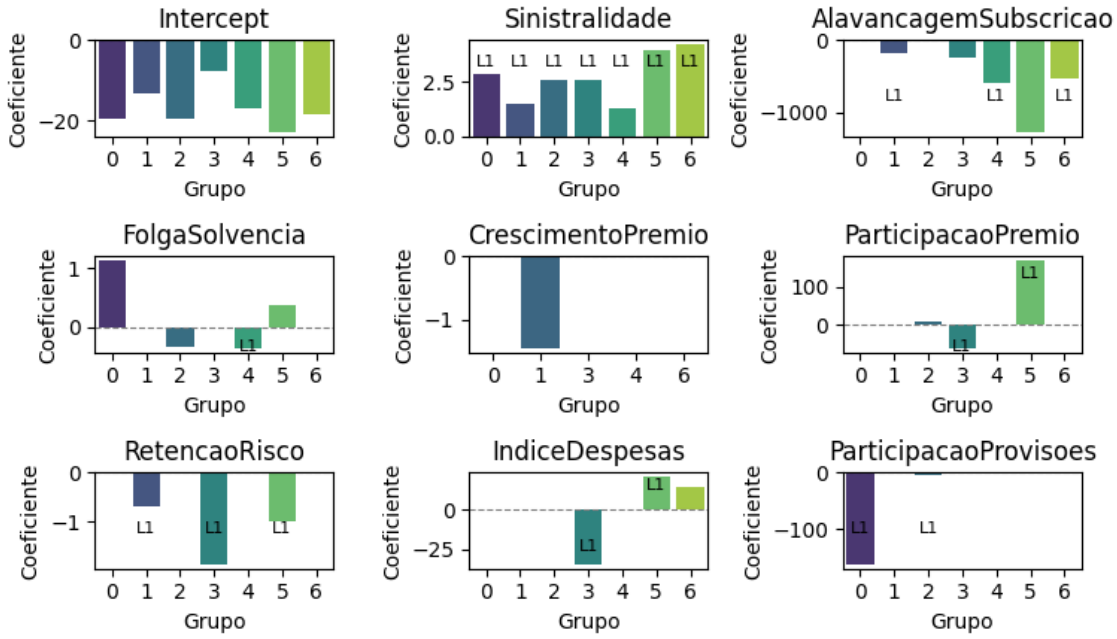
Fonte: Elaboração própria.

Os dados de diagnóstico dos modelos lineares generalizados para cada agrupamento indicam autocorrelação nos resíduos, conforme teste de Ljung-Box, exceto nos grupos 4 e cinco, e presença de heterogeneidade na variância dos resíduos para os agrupamentos 0 e 5, que indica que pode haver outros determinantes da sinistralidade não incluídos no modelo ou variância funcional sub-especificada.

Em razão do diagnóstico, o erro-padrão dos coeficientes foi calculado utilizando um estimador de matriz de covariância robusto à heterocedasticidade, HC3. A escolha da variante

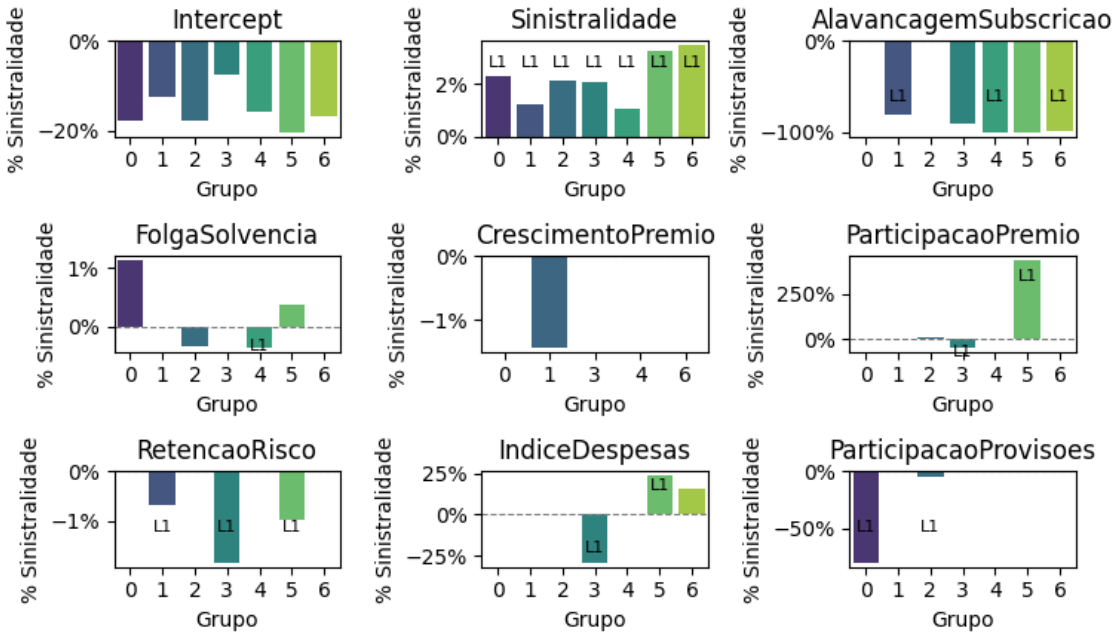
HC3 dentre os estimadores Heterocedasticidade-Consistentes justifica-se por ser o que apresenta melhor desempenho em amostras pequenas e influentes, devido ao desconto do quadrado do peso que a alavancagem (*leverage*) das observações possa exercer sobre a variância dos resíduos (Chibari-Neto; Galvão, 2003).

Figura 14. Comparativo entre coeficientes significativos através dos grupos de mercado



Fonte: Elaboração própria.

Figura 15. Comparativo entre efeitos percentuais significativos através dos grupos de mercado



O estimador robusto para o erros-padrão permite testes estatísticos confiáveis mesmo quando a suposição de variância constante nos resíduos é violada, de modo que é possível analisar a intensidade dos coeficientes nos modelos ajustados, conforme encontra-se consolidada na Figura 14.

Observa-se que todos os modelos apresentam coeficiente significativo para a sinistralidade defasada (L1), enquanto que as demais variáveis derivadas no estudo são aproveitadas por pelo menos um de cada agrupamento de mercado.

Em razão do efeito multiplicativo exponencial da função de estimação log-link, os fatores observáveis de mercado que exercem maior influência através dos grupos segmentados são: Alavancagem de Subscrição, Folga de Solvência e Retenção de Riscos. A Figura 15 consolida os efeitos estatisticamente significativos sobre a variação da Sinistralidade, em percentual, dado o incremento de 1 ponto percentual nas variáveis do estudo.

Nos agrupamentos 1, 3 e 5, o aumento de 0,01 na proporção de riscos retidos do ramo no período passado reduz a sinistralidade observada entre 1% e 2%. Fatores de escala operacional das companhias podem conferir maior gestão sobre os riscos retidos, notadamente em mercados de especialização como o do ramo 0351.

A Folga de Solvência se mostra um fator com efeito significativo em 4 grupos do mercado, de modo que o aumento em 0,01 nessa distância entre o patrimônio ajustado e o limite regulatório das companhias gera efeito tanto negativo como positivo sobre a sinistralidade observada, da ordem de ±1% a depender do agrupamento de entidades estudado. Enquanto níveis mais elevados de capital podem aumentar a capacidade de absorção e perdas (Cummins; Sommer, 1996), também podem estar associados a ineficiências na alocação de recursos e problemas de agência, resultando em um pior desempenho operacional.

Por fim, a medida de exposição a riscos do ramo em fração do patrimônio ajustado, denominada Alavancagem de Subscrição, exerce efeito negativo sobre a sinistralidade em cinco dos sete agrupamentos de companhias estudados. Merece ressalvar que o aumento de 1 ponto percentual no domínio da série dessa variável já representa uma grande alteração nos valores observados, o que, por sua vez, também é um achado condizente com o resultado. Observa-se que no estado atual do mercado o aumento na exposição ao risco em relação ao capital disponível pode estar associado a um melhor aproveitamento da capacidade operacional das seguradoras.

Os modelos de agrupamento permitem concluir que a dinâmica da sinistralidade não é uniforme no mercado, visto que apresenta grupos de seguradoras com perfis distintos de

operação. Ademais, a relação entre a sinistralidade média e as variáveis explicativas varia a depender de tais agrupamentos.

5 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo investigar quais fatores observáveis do mercado segurador explicam a dinâmica da sinistralidade no ramo 0351 – Seguro de Responsabilidade Civil Geral – no Brasil. Para isso, foi conduzida análise empírica baseada em dados mensais reportados pelas seguradoras à Superintendência de Seguros Privados (SUSEP), abrangendo o período de novembro de 2016 a maio de 2025. A investigação combinou diferentes abordagens quantitativas, incluindo modelagem de séries temporais, modelos multivariados, modelagem estatística por meio de modelos lineares generalizados e técnicas de agrupamento para segmentação do mercado.

Inicialmente, a análise da dinâmica temporal da sinistralidade indicou que a série apresenta persistência e padrões sazonais estatisticamente relevantes. Esse resultado sugere que o comportamento da sinistralidade pode ser analisado de maneira sistemática a partir de informações observáveis do próprio mercado segurador.

Na sequência, a modelagem multivariada por meio de um modelo VAR permitiu examinar a relação entre a sinistralidade e diferentes variáveis derivadas da estrutura operacional das seguradoras. Os resultados indicaram associação estatisticamente significativa entre a sinistralidade e variáveis como alavancagem de subscrição, participação em provisões e índice de despesas, além da própria defasagem da sinistralidade. Os achados sugerem que o comportamento das perdas securitárias está relacionado tanto à dinâmica interna das operações das seguradoras quanto às condições estruturais do mercado.

A modelagem estatística por meio de modelos lineares generalizados possibilitou avaliar quais fatores explicam estatisticamente os níveis de sinistralidade observados. Os resultados indicaram que variáveis como participação no prêmio, participação em provisões, retenção de riscos e a própria defasagem da sinistralidade apresentam associação estatisticamente significativa com a sinistralidade média do mercado.

Por fim, a análise de agrupamento evidenciou que o mercado do ramo 0351 apresenta estrutura heterogênea. A segmentação das seguradoras em diferentes grupos revelou que a dinâmica da sinistralidade não é uniforme entre os participantes do mercado, sendo influenciada por diferentes estratégias operacionais, níveis de exposição ao risco e estruturas de capital. A aplicação de modelo GLM específico para cada agrupamento reforçou esse diagnóstico, mostrando que a intensidade e a direção do efeito das variáveis explicativas variam entre os diferentes segmentos de mercado.

Dessa forma, os resultados obtidos permitem responder ao problema de pesquisa proposto. A dinâmica da sinistralidade no ramo 0351 pode ser explicada por um conjunto de fatores observáveis relacionados à estrutura operacional das seguradoras, em especial aqueles associados à retenção de riscos, à constituição de provisões técnicas, à estrutura de capital e à própria persistência temporal da sinistralidade. Esses fatores refletem tanto decisões estratégicas das seguradoras quanto condicionantes institucionais do mercado segurador.

Do ponto de vista teórico, o estudo contribui ao integrar diferentes abordagens econométricas e estatísticas na análise da sinistralidade em seguros de responsabilidade civil no Brasil, campo ainda relativamente pouco explorado na literatura acadêmica nacional. A combinação entre modelos de séries temporais, modelagem estatística e técnicas de agrupamento permitiu examinar o fenômeno sob múltiplas perspectivas, oferecendo evidências empíricas sobre os determinantes da sinistralidade nesse ramo específico do mercado segurador.

Sob a perspectiva prática, os resultados obtidos podem contribuir para o aprimoramento da gestão de riscos e da subscrição no mercado segurador. A identificação de fatores associados à sinistralidade pode auxiliar seguradoras, reguladores e analistas de mercado na compreensão das dinâmicas operacionais que influenciam o desempenho técnico das companhias, bem como no desenvolvimento de estratégias de gestão de capital e transferência de riscos mais alinhadas ao comportamento observado do mercado.

Entretanto, o estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. A análise foi conduzida com base em dados agregados reportados pelas seguradoras, o que limita a observação de características mais detalhadas das apólices e dos riscos segurados. Além disso, a presença de heterogeneidade estrutural entre as seguradoras sugere que outros fatores não observados podem influenciar a dinâmica da sinistralidade, os quais não foram capturados pelas variáveis consideradas no modelo.

Diante dessas limitações, pesquisas futuras podem ampliar a análise incorporando informações adicionais sobre características dos contratos no ramo por região, ou inclusão de indicadores macroeconômicos nos modelos. Também podem ser exploradas abordagens alternativas de modelagem, incluindo técnicas de aprendizado de máquina, capazes de capturar de forma mais direta as heterogeneidades presentes no mercado segurador.

Em síntese, os resultados deste estudo indicam que a sinistralidade no ramo 0351 apresenta dinâmica estruturada e associada a fatores observáveis do mercado segurador, reforçando a importância da análise quantitativa para compreensão e gestão dos riscos no setor de seguros.

REFERÊNCIAS

ALVIM, P. **O contrato de seguro.** Rio de Janeiro: Forense. 2001.

BAKER, T. The law and economics of liability insurance. In: FAURE, M. (ed.). **Trust and insurance contracts**. 2. ed. Cheltenham: Edward Elgar, 2013. p. 23-60.

BARNETT, V; LEWIS, T. **Outliers in statistical data**. 3 ed. New York: Wiley, 1994.

BELSEY, D. A., KUH, E.; WELSCH, R. E. **Regression diagnostics: identifying influence data and source of collinearity**. New York: Wiley, 1980.

BOX, G. E. P.; JENKINS, G. M; Reinsel, G. C; Ljung, G. M. **Time Series Analysis**: forecasting and control. 5 ed. Hoboken, New Jersey: John Wiley & Sons, 2016.

BRASIL. **Lei nº 10.406, de 10 de janeiro de 2002**. Institui o Código Civil. Diário Oficial da União: seção 1, Brasília, DF, ano 139, n. 7, 11 jan. 2002.

CHIBARI-NETO, F.; GALVÃO, N. M. S. A class of improved heteroskedasticity-consistent covariance matrix estimators. **Communications in Statistics - Theory and Methods**, v. 32(10). 2003.

CLEVELAND, R. B.; CLEVELAND, W. S.; MCRAE, J. E.; TERPENNING, I. STL: A seasonal-trend decomposition procedure based on Loess. **Journal of Official Statistics**, v. 6, n. 1, 1990.

CNSEG. **Conjuntura CNSEG**: editorial e análise de mercado. Ano 8. nº 122. Julho 2025.

CNSP. **Resolução nº 432, de 12 de novembro de 2021**. Dispõe sobre provisões técnicas, ativos redutores da necessidade de cobertura das provisões técnicas, capitais de risco, patrimônio líquido ajustado, capital mínimo requerido, planos de regularização, limite de retenção, critérios para a realização de investimentos, normas contábeis, auditoria contábil e auditoria atuarial independentes e Comitê de Auditoria aplicáveis a sociedades seguradoras, entidades abertas de previdência complementar, sociedades de capitalização e resseguradores.

CNSP. **Resolução nº 407, de 30 de março de 2021**. Dispõe sobre a classificação dos ramos de seguros.

COASE, R. H. The problem of social cost. **Journal of Law and Economics**, v. 3, p. 1-44, 1960

COELHO, F. U. **Manual de direito comercial**: direito de empresa. São Paulo: Revista dos Tribunais, 2016.

CUMMINS, J. D.; OUTREVILLE, J. F. An international analysis of underwriting cycles in property-liability insurance. **Journal of Risk and Insurance**, v. 54, n. 2. 1987.

CUMMINS , J. D.; SOMMER, D. W. Capital and risk in property-liability insurance markets. **Journal of Banking & Finance**, v. 20, n. 6, 1996.

CUMMINS, J. D.; RUBIO-MISAS, M. Deregulation, consolidation and efficiency: evidence from the Spanish insurance industry. **Journal of Money, Credit and Banking**, v. 38, n. 2, p. 323-355. 2006.

CUMMINS, J. D.; PHILLIPS, R. D. Capital adequacy and insurance risk-based capital systems. **Journal of Insurance Regulation**, v. 28, n. 1, p. 323-355. 2009.

DICKEY, D. A; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root in **Journal of the American Statistical Association**, v. 74, n. 366. 1979.

DOBSON, A. J.; BARNETT, A. **An Introduction to Generalized Linear Models**. 4. ed. Boca Raton: CRC Press, 2018.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. Chichester: Willey, 2011.

FERREIRA, Pedro G. C. **Análise de Séries Temporais em R**: Curso Introdutório.Rio de Janeiro: Elsevier-FGV IBRE, 2018.

FREES, E. W.; DERRIG, R. A.; MEYERS, G. **Predictive Modeling Applications in Actuarial Science**, Volume I. Cambridge: Cambrige University Press, 2014.

GAGLIANO, Pablo S.; PAMPLONA FILHO, Rodolfo. **Novo curso de direito civil**: responsabilidade civil.São Paulo: Saraiva, 2009.

GAN, G.; VALDEZ, E.; HUANG, H. **Model-based cluster analysis for insurance claims**. North American Actuarial Journal, v. 24, n. 3, 2020.

GIL, A. C**. Como elaborar projetos de pesquisa**. 4 ed. Atlas. São Paulo, 2002.

GUJARATI, Damodar N. **Econometria Básica**. Porto alegre: AMGH, 2011.

HAMILTON, J. D. **Time Series Analysis**. Princeton: Princeton University Press, 1994.

JORGENSEN, B. Exponential dispersion models. **Journal of the Royal Statistical Society**: **Series B (Methodological),** n. 2, 1987.

KELEJIAN, H. H. E ROBINSON, D. A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding Monte Carlo results. **Regional Science and Urban Economics**, vol. 28, 1998.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. Em: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. Berkley: Universidy of California Press, 1967. v. 1..

MARCONI, M. A; LAKATOS, E. M. **Fundamentos de metodologia científica**. 8 ed. São Paulo: Atlas, 2017.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society**, v. 135, n. 3, 1972.

OLIVEIRA, L. M. de; TAVARES, R. C. Transformações no mercado de resseguro de responsabilidade civil geral no Brasil. **Revista eletrônica do departamento de ciências contábeis & atuária e métodos quantitativos – Redeca**. v. 1. n. 2. 2014.

PANDOLFI, A. S.; GONÇALVES, J. N. Previsão da sinistralidade em seguros de vida utilizando modelos de séries temporais. **Revista ENIAC Pesquisa**. v. 13. n. 1. 2024.

PEREIRA, C. M. da S. **Instituições de direito civil**: introdução ao direito civil; teoria geral de direito civil. Forense, Rio de Janeiro: Forense, 2004.

PRADO, C. A. **Perspectivas para os seguros de responsabilidade civil**. 2021.

ROA, A. D.; GONSALVES, R. A. Modelos probabilísticos de severidade para grandes perdas. **Revista eletrônica do departamento de ciências contábeis & atuária e métodos quantitativos – Redeca**. v. 2. n. 1. 2015.

ROUSSEEUW, P, J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**. v. 20, 1986.

SANDSTRÖM, A. **Handbook of Solvency for Actuaries and Risk Managers**. Boca Raton: CRC Press, 2011.

SHI, P.; SHI, K. Nonlife insurance risk classification using categorical embedding. **North American Actuarial Journal**, v. 26, n. 2. 2022.

SHIM, J. Capital-based regulation and insurer risk-taking. **Journal of Risk and Insurance**, v. 77, n. 4, 2010.

SIMS, C. A. **Macroeconomics and reality**. Econometrica, v. 48, n. 1, 1980.

SUSEP. **Circular nº 437, de 5 de julho de 2012**. Dispõe sobre normas gerais de operação do seguro de responsabilidade civil geral.

SUSEP. **Circular nº 517, de 30 de julho de 2015**. Dispõe sobre provisões técnicas; teste de adequação de passivos; ativos redutores; capital de risco de subscrição, crédito, operacional e mercado; constituição de banco de dados de perdas operacionais; plano de regularização de solvência; registro, custódia e movimentação de ativos, títulos e valores mobiliários garantidores das provisões técnicas; Formulário de Informações Periódicas - FIP/SUSEP; Normas Contábeis e auditoria contábil independente das seguradoras, entidades abertas de previdência complementar, sociedades de capitalização e resseguradores; exame de certificação e educação profissional continuada do auditor contábil independente e sobre os Pronunciamentos Técnicos elaborados pelo Instituto Brasileiro de Atuária - IBA.

SUSEP. **Circular nº 517, de 27 de julho de 2021**. Dispõe sobre os seguros do grupo responsabilidades.

SUSEP. **Relatórios de Estabilidade e Solvência**. 2022.

VAUGHAN, E. J.; VAUGHAN, **T. M. Fundamentals of Risk and Insurance**. 11. ed. Hoboken: Wiley, 2014

WILKINSON, L. **The Grammar of Graphics**. 1 ed. New York: Springer-Verlag, 2005.

APÊNDICE A – SCRIPT DA MODELAGEM EMPÍRICA

```
# bibliotecas e configurações gerais
import pandas as pd
import numpy as np
np.set_printoptions(suppress=True, precision=2)
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm

import warnings
warnings.filterwarnings('ignore')

%%capture
# caminho da base de dados primária
# import requests
from zipfile import ZipFile
from google.colab import drive
drive.mount('/content/drive')

# zip_url = 'https://www2.susep.gov.br/redarq.asp?arq=BaseCompleta%2ezip'
zip_path = '/content/drive/MyDrive/TCC-dados/BaseCompleta.zip'

# Baixar o arquivo ZIP
# print(f"Downloading {zip_url}...")
# response = requests.get(zip_url)
# response.raise_for_status()
# with open(zip_path, 'wb') as f:
#     f.write(response.content)
# print(f"File downloaded to {zip_path}")
```

```
# fontes adotadas para construção da base de pesquisa
base_com_ramo = ['Ses_seguros.csv', 'ses_provramos.csv']
base_sem_ramo = ['Ses_pl_margem.csv', 'Ses_seg_prov_det.csv']

# função de coleta de dados da base primária
def coleta(nome, ramo = True):
  with ZipFile(zip_path, 'r') as zip_object:
    dados = pd.read_csv(zip_object.open(nome), sep=';', decimal=',', low_memory=False)
    if ramo==True:
      dados = dados[(dados['coramo']==351) & (dados['damesano']>=201612) & (dados['damesano']<=202505)]
    else:
      dados = dados[(dados['damesano']>=201612) & (dados['damesano']<=202505)]
    dados.sort_values(by=['damesano'], inplace=True)
    dados['ID'] = dados['damesano'].astype(str)+'_'+dados['coenti'].astype(str)
  return dados

# coleta dos dados atrelados ao ramo 0351 (prêmio, sinistro, desp. comercial, etc.)
dados_com_ramo = [coleta(nome) for nome in base_com_ramo]
dados_com_ramo[0] = dados_com_ramo[0].iloc[:,[21,0,1,2,5,7,13,15,10,14]]
dados_com_ramo[1] = dados_com_ramo[1].iloc[:,[22,3,7,8]]
dados_com_ramo = pd.merge(dados_com_ramo[0], dados_com_ramo[1], on='ID', how='outer')

# coleta dos dados não atrelados ao ramo (pla ajustado, cmr, provisoes, etc.)
dados_sem_ramo = [coleta(nome, ramo=False) for nome in base_sem_ramo]
dados_sem_ramo[0] = dados_sem_ramo[0].iloc[:,[9,7,8]]
dados_sem_ramo[1] = dados_sem_ramo[1].iloc[:,[23,2,10,12]]
dados_sem_ramo = pd.merge(dados_sem_ramo[0], dados_sem_ramo[1], on='ID', how='outer')

# consolidação da base de pesquisa
dados = pd.merge(dados_com_ramo, dados_sem_ramo, on='ID', how='left')
```

```
# remoção de informação base nula
zero = dados.groupby('coenti')['premio_ganho'].sum()[
    dados.groupby('coenti')['premio_ganho'].sum() > 0].index
dados = dados[dados['coenti'].isin(zero)]

# remoção de informação base vazia
for i in list(set(dados['coenti'])):
  if dados[dados['coenti']==i].shape[0]<102:
    dados.drop(index=dados[dados['coenti']==i].index, inplace=True)

# conversão de `damesano` em formato data
dados['damesano'] = dados['damesano'].astype(int).astype(str)
dados['damesano'] = pd.to_datetime(dados['damesano'], format='%Y%m')

# agregação de provisões
dados['provisoes_0351'] = dados[['ppng', 'psl', 'ibnr']].sum(axis=1)
dados['provisoes_totais'] = dados[['prem_n_ganhos', 'sinistros_liquidar',
                                   'sinistros_ibnr']].sum(axis=1)

## inclusão do premio ganho total no periodo para todos os ramos
premio_ganho_total = coleta('Ses_seguros.csv', False)
premio_ganho_total.loc[premio_ganho_total['premio_ganho']<0, 'premio_ganho'] = 0
premio_ganho_total = premio_ganho_total.groupby('ID')['premio_ganho'].sum().reset_index()
premio_ganho_total.rename(columns={'premio_ganho': 'premio_ganho_total'}, inplace=True)
dados = pd.merge(dados, premio_ganho_total, on='ID', how='left')

## inclusão do risco retido no ramo 0351
dados['risco_retido'] = (dados['premio_de_seguros'] - dados['despesa_resseguros']
                         + dados['receita_resseguro'])

## descarte de variáveis já combinadas
```

```
dados = dados.drop(columns=['ppng', 'psl', 'ibnr',
                            'prem_n_ganhos', 'sinistros_liquidar','sinistros_ibnr',
                            'despesa_resseguros', 'receita_resseguro'])

# lista de entidades abrangidas
with ZipFile(zip_path, 'r') as zip_object:
  cias = pd.read_csv(zip_object.open('Ses_cias.csv'), sep=';', decimal=',',
                     low_memory=False, encoding='latin1')
seg = list(set(dados['coenti']))
nomes = [cias[cias['Coenti']==coenti]['Noenti'].values[0] for coenti in seg]
entidades = pd.DataFrame({'código': seg, 'entidade': nomes}).sort_values(by='entidade')
print(entidades)

# criaçao de variáveis acumuladas em 12 meses
for c in dados.columns[4:]:
  dados[f'{c}_12m'] = dados.groupby('coenti')[c].transform(
      lambda x: x.rolling(window=12, min_periods=1).sum())

# coleta do IPCA mensal
%%capture
!pip install sidrapy
import sidrapy

def coleta_ipca(periodo):
  indice = sidrapy.get_table(table_code="1737", territorial_level="1",
                             ibge_territorial_code="all",
                             variable='63', period=periodo)
  return indice.V[1]

IPCAm = [float(coleta_ipca(i)) for i in dados.damesano.dt.strftime('%Y%m').unique()]
```

```
# inflação das séries acumuladas para cada entidade
def inflator(v):
  acum = np.cumprod(1+(np.array(v)/100))
  return acum/acum[-1]

for c in list(dados.columns[14:24]):
  dados[f'{c}_inf'] = np.zeros(dados.shape[0])
  for coenti in seg:
    dados.loc[dados['coenti']==coenti, f'{c}_inf'] = dados.loc[dados['coenti']==coenti, c]/inflator(IPCAm)
# base média do mercado acumulado em 12 meses no ramo 0351
df_12m = (dados[['damesano'] + [c for c in dados.columns if '_12m' in c]]
          .groupby('damesano').mean().reset_index().iloc[12:])

# análise gráfica das variáveis acumladas em 12 meses
fig, eixo = plt.subplots(5, 2, figsize=(6.5,9.5))
for i, c in enumerate(list(df_12m.columns[1:11])):
  ax = eixo[i//2, i%2]
  g1 = sns.lineplot(data=df_12m, x='damesano', y=c, ax=ax, label='obs')
  g2 = sns.lineplot(data=df_12m, x='damesano', y=f'{c}_inf', ax=ax, label='inf')
  ax.set_xlabel('Ano')
  ax.set_ylabel('')
  ax.set_title(c)
  plt.setp(ax.get_xticklabels(), rotation=45, ha='right')
  plt.tight_layout()
plt.tight_layout()

# gerar de taxas
def gerar_taxas(dados=dados):
  # sinistralidade
  dados['Sinistralidade'] = dados['sinistro_ocorrido_12m_inf']/dados['premio_ganho_12m_inf']
  # índice de despesas
```

```
  dados['IndiceDespesas'] = dados['desp_com_12m_inf']/dados['premio_ganho_12m_inf']
  # participação do ramo nas provisões
  dados['ParticipacaoProvisoes'] = dados['provisoes_0351_12m_inf']/dados['provisoes_totais_12m_inf']
  # concentração de premio no ramo
  dados['ParticipacaoPremio'] = dados['premio_ganho_12m_inf']/dados['premio_ganho_total_12m_inf']
  # folga de solvência
  dados['FolgaSolvencia'] = dados['NovoPla_12m_inf']/dados['CMR_12m_inf']
  # retenção de riscos
  dados['RetencaoRisco'] = dados['risco_retido_12m_inf']/dados['premio_ganho_12m_inf']
  # crescimento do prêmio
  dados['CrescimentoPremio'] = dados.groupby('coenti')['premio_de_seguros_12m_inf'].pct_change()
  # avalancagem de subscrição
  dados['AlavancagemSubscricao'] = dados['premio_ganho_12m_inf']/dados['NovoPla_12m_inf']

  # remover indeterminações
  taxas = list(dados.columns[34:])
  dados.loc[:, taxas] = dados.loc[:, taxas].fillna(0).replace([np.inf, -np.inf], 0)
  return dados, taxas

# formação da base de taxas médias no período
dados, taxas = gerar_taxas()
df_taxas = dados[['damesano'] + taxas].groupby('damesano').mean().reset_index().iloc[12:]
df_taxas.set_index('damesano', inplace=True)

# análise grafica de comportamento aberrante em Sinistralidade
fig, eixo = plt.subplots(4, 2, figsize=(6.5, 9.5))
fig.delaxes(eixo[3,1])
for i in range(7):
  coenti = list(dados.groupby('coenti')['premio_de_seguros_12m'].max().reset_index( ).
                sort_values(by='premio_de_seguros_12m', ascending=False).
                iloc[range(i*6,(i+1)*6), :].iloc[:,0])
```

```
  ax = eixo[i//2, i%2]
  for coenti in coenti:
    dados[dados['coenti'] == coenti].iloc[12:].plot(
        x='damesano', y='Sinistralidade', ax=ax, label=coenti)
  ax.set_xlabel('Ano')
  ax.set_ylabel('Sinistralidade')
  ax.legend(loc='upper right', ncol=2, fontsize='small')
  ax.grid(True)
  ax.set_title(f'Top {i*6+1}-{min((i+1)*6, 42)} Entidades')
  plt.setp(ax.get_xticklabels(), rotation=45, ha='right')
plt.tight_layout()
plt.show()

# analise de premios e sinistros para as entidades com sinistralidade aberrante
aberr = [6921, 3727, 5193, 5941, 5045, 6793]
fig, eixo = plt.subplots(3, 2, figsize=(6.5, 7.5))
for i, coenti in enumerate(aberr):
  ax = eixo[i//2, i%2]
  dados[dados['coenti'] == coenti].iloc[12:].plot(
      x='damesano', y=['premio_ganho_12m', 'sinistro_ocorrido_12m'], ax=ax)
  ax.set_title(f'Entidade {coenti}')
  ax.set_xlabel('Ano')
  ax.grid(True)
  plt.setp(ax.get_xticklabels(), rotation=45, ha='right')
plt.tight_layout()
plt.show()

# refazimento da base de dados
dados.drop(index=dados[dados['coenti'].isin(aberr)].index, inplace=True)

# resetar taxas
```

```
dados, taxas = gerar_taxas()
df_taxas = dados[['damesano'] + taxas].groupby('damesano').mean().reset_index().iloc[12:]
df_taxas.set_index('damesano', inplace=True)

# analise grafica de comportamento aberrante ao longo das taxas
df_taxas_coenti = (dados.groupby('coenti').apply(lambda x: x[taxas].iloc[12:].
                                                 mean()).reset_index())

fig, eixo = plt.subplots(4, 2, figsize=(6.5, 9.5))
fig.delaxes(eixo[3,1])

for i, c in enumerate(taxas[1:]):
  eixo[i//2, i%2].scatter(df_taxas_coenti['Sinistralidade'], df_taxas_coenti[c],
                          cmap='viridis', alpha=0.7)
  for j, coenti_name in enumerate(df_taxas_coenti['coenti']):
    eixo[i//2, i%2].text(df_taxas_coenti['Sinistralidade'][j], df_taxas_coenti[c][j],
                         str(coenti_name), fontsize=6, ha='right')

  eixo[i//2, i%2].set_xlabel('Sinistralidade')
  eixo[i//2, i%2].set_ylabel(c)
  eixo[i//2, i%2].grid(True)

plt.tight_layout()
plt.show()

# entidades com comportamento aberrante ao longo das taxas
aberr = [6785, 5908, 5118, 2461, 3069]

# refazimento da base de dados
dados.drop(index=dados[dados['coenti'].isin(aberr)].index, inplace=True)
```

```
# resetar taxas
dados, taxas = gerar_taxas()
df_taxas = dados[['damesano'] + taxas].groupby('damesano').mean().reset_index().iloc[12:]
df_taxas.set_index('damesano', inplace=True)

# analise grafica das taxas médias - boxplot
fig, eixo = plt.subplots(4, 2, figsize=(5.5,5.5))
[sns.boxplot(x=df_taxas[c], ax=eixo[i//2, i%2]) for i, c in enumerate(taxas)]
plt.tight_layout()

# limitar extremos nas taxas médias
def winsor(c, q, sup=True, inf=True):
  for c in c:
    if sup==True:
      df_taxas.loc[df_taxas[c] >= df_taxas[c].quantile(q), c] = df_taxas[c].quantile(0.75)
    if inf==True:
      df_taxas.loc[df_taxas[c] <= df_taxas[c].quantile(1-q), c] = df_taxas[c].quantile(0.25)

# limitar extremos em IndiceDespesas
winsor(['IndiceDespesas'], 0.99, sup=False)

# diagnóstico de endogeneidade
import statsmodels.api as sm
from statsmodels.formula.api import ols

res_endog = []
f_test_iv = []
preditores = [p for p in taxas if p != 'Sinistralidade']

for taxa in preditores:
  df_endog = df_taxas.copy()
```

```
    df_endog[f'{taxa}_lag1'] = df_endog[taxa].shift(1)
    df_endog.dropna(inplace=True)

    endog = taxa
    exog = [p for p in preditores if p != taxa]
    instr = f'{taxa}_lag1'
    df_endog = sm.add_constant(df_endog, prepend=False)

    formula1 = f'{endog} ~ {'+'.join([c for c in (['const'] + exog + [instr])])}'
    reg1 = ols(formula1, data=df_endog).fit()
    df_endog['v_hat'] = reg1.resid
    #f_test_iv.append({'taxa': taxa, 'p-valor': reg1.f_test([

    formula2 = f'Sinistralidade ~ {'+'.join([c for c in (['const'] + exog + ['v_hat'])])}'
    reg2 = ols(formula2, data=df_endog).fit()

    res_endog.append({'taxas': taxa, 'p-valor': reg2.pvalues['v_hat'],
                      'relevancia_IV': reg1.f_test(f'{instr} = 0').fvalue})

print(pd.DataFrame(res_endog).round(4).to_string())

# transformação das variáveis endogenas em instrumentos
l1 = ['ParticipacaoProvisoes', 'FolgaSolvencia', 'RetencaoRisco', 'AlavancagemSubscricao']
for c in l1: df_taxas[f'{c}_l1'] = df_taxas[c].shift(1)
df_taxas.dropna(inplace=True)

# análise de estacionariedade
from statsmodels.tsa.stattools import adfuller
def ADF(serie):
    df_adf = []
    for c in serie.columns:
```

```
    adf = adfuller(serie[c], autolag='AIC')
    df_adf.append({'taxa': c, 'tau': adf[0], 'p-valor': adf[1], 'lags': adf[2],
                   '1%': adf[4]['1%'],
                   '5%': adf[4]['5%'],
                   '10%': adf[4]['10%']})
  print(pd.DataFrame(df_adf).set_index('taxa').round(4).to_string())

ADF(df_taxas)

# Diferenciação de variáveis não estacionárias
d1 = ['IndiceDespesas', 'ParticipacaoPremio', 'RetencaoRisco_l1', 'AlavancagemSubscricao_l1']
for c in d1: df_taxas[f'{c}_d1'] = df_taxas[c].diff(1)
df_taxas.dropna(inplace=True)

# decomposição de 'ParticipacaoPremio' (detrending)
from statsmodels.tsa.seasonal import STL
df_taxas['ParticipacaoPremio_dt'] = STL(df_taxas.ParticipacaoPremio, period=12, robust=True).fit().resid

ADF(df_taxas)

# analise de normalidade - histogramas
fig, eixo = plt.subplots(4, 2, figsize=(6.5,7.5))
for i, c in enumerate(df_taxas.columns):
  ax = eixo[i//2, i%2]
  g = sns.histplot(data=df_taxas, x=c, kde=True, ax=ax)
  ax.set_xlabel(c)
  ax.set_ylabel('')
plt.tight_layout()

# analise de normalidade - qq-plot
fig, eixo = plt.subplots(4, 2, figsize=(6.5,7.5))
```

```
for i, c in enumerate(df_taxas.columns):
  ax = eixo[i//2, i%2]
  g = sm.qqplot(df_taxas[c], line='s', ax=ax)
  ax.set_xlabel('')
  ax.set_ylabel('')
  ax.set_title(c)
plt.tight_layout()

# matriz de correlação das taxas
plt.subplots(figsize=(5.5,5.5))
corr = df_taxas.corr().round(2)
mask = np.triu(np.ones_like(corr, dtype=bool))
sns.heatmap(corr, mask=mask, annot=True, cmap='coolwarm', center=0, vmin=-1, vmax=1)

# teste do fator de incremento da variância
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant

X = df_taxas.drop(columns=['Sinistralidade'])
X = add_constant(X)
fiv = pd.DataFrame()
fiv['taxa'] = X.columns
fiv['FIV'] = [variance_inflation_factor(X.values, i) for i in range(8)]

print("Fator de Incremento da Variância:")
print(fiv.round(2).to_string())

# MODELO SARIMA

%%capture
# biblioteca
```

```
from statsmodels.tsa.statespace.sarimax import SARIMAX

!pip install pmdarima
from pmdarima import auto_arima
auto_arima(df_taxas.Sinistralidade, start_p=1, start_q=1, vtest='adf', m=12, seasonal=True, trace=True)
sarima = SARIMAX(df_taxas.Sinistralidade, order=(2,0,1),seasonal_order=(2,0,0,12))
res_sarima = sarima.fit()

# comparação entre valores observados e ajustados no modelo SARIMA
plt.subplots(figsize=(5.5, 3.0))
df_taxas.Sinistralidade.plot(label='Sinistralidade observada', lw=1, style='k--')
res_sarima.fittedvalues.plot(label='Sinistralidade SARIMA')
res_sarima.resid.plot(label='residuos')
plt.xlabel('')
plt.legend()
plt.tight_layout()
plt.show()

# MODELO VAR

# biblioteca
from statsmodels.tsa.api import VAR
from statsmodels.tsa.stattools import grangercausalitytests
from statsmodels.stats.stattools import durbin_watson

# Modelo
var = VAR(df_taxas)
# critérios de selecão de lag
print(var.select_order().summary())

# resultados do modelo VAR
```

```
res_var = var.fit(maxlags=8, ic='aic')

# comparação entre valores observados e ajustados no modelo VAR
plt.subplots(figsize=(5.5, 3.0))
df_taxas['Sinistralidade'].plot(label='Sinistralidade observada', lw=1, style='k--')
res_var.fittedvalues.Sinistralidade.plot(label='Sinistralidade VAR')
res_var.resid.Sinistralidade.plot(label='residuos')
plt.xlabel('')
plt.legend()
plt.tight_layout()
plt.show()

# análise de granger-causalidade entre as taxas
df_granger = pd.DataFrame()
for c in df_taxas.columns:
  for l in df_taxas.columns:
    res = grangercausalitytests(df_taxas[[c, l]], maxlag=8, verbose=False)
    p_valor = [round(res[i+1][0]['ssr_ftest'][1], 4) for i in range(8)]
    df_granger.loc[c, l] = min(p_valor)
df_granger.columns = [x + '_x' for x in df_taxas.columns]
df_granger.index = [y + '_y' for y in df_taxas.columns]

# matriz de granger-causalidade
plt.subplots(figsize=(6.5,5.5))
sns.heatmap(df_granger, annot=True, fmt='.2f', cmap='coolwarm', center=0.05)
plt.suptitle('Matriz de Granger-Causalidade')
plt.title(r'p-valores ($\alpha$=0.05)')
plt.tight_layout()
plt.show()

# MODELO GLM
```

```
# biblioteca
import statsmodels.api as sm
import statsmodels.formula.api as smf

# base do modelo
df_glm = dados.groupby('damesano')[taxas].mean().iloc[12:]
df_glm['logPremioEmitido'] = np.log(dados.groupby('damesano')['premio_de_seguros_12m_inf'].
                                    mean().iloc[12:])

# inclusão de componente autorregressivo da sinistralidade
def autorr_sinist(df_): df_['Sinistralidade_l1'] = df_['Sinistralidade'].shift(1)
autorr_sinist(df_glm)

# diagnostico e tratamento de endogeneidade
def diag_endog(df_, variaveis_):
  res_endog = []
  f_test_iv = []
  preditores = [p for p in variaveis_ if p != 'Sinistralidade']

  for taxa in preditores:
    df_endog = df_.copy()
    df_endog[f'{taxa}_lag1'] = df_endog[taxa].shift(1)
    df_endog.dropna(inplace=True)

    endog = taxa
    exog = [p for p in preditores if p != taxa]
    instr = f'{taxa}_lag1'
    df_endog = sm.add_constant(df_endog, prepend=False)

    # 1o estágio
```

```
    formula1 = f'{endog} ~ {'+'.join([c for c in (['const'] + exog + [instr])])}'
    reg1 = ols(formula1, data=df_endog).fit()
    df_endog['v_hat'] = reg1.resid

    # 2o estágio
    formula2 = f'Sinistralidade ~ {'+'.join([c for c in (['const'] + exog + ['v_hat'])])}'
    reg2 = ols(formula2, data=df_endog).fit()

    res_endog.append({'taxas': taxa, 'p-valor': reg2.pvalues['v_hat'],
                      'relevancia_IV': reg1.f_test(f'{instr} = 0').fvalue})

  return pd.DataFrame(res_endog)

def trat_endog(diag, df_):
  l1 = list(diag[(diag['p-valor'] <= 0.05) & (diag['relevancia_IV']>10)]['taxas'])
  for taxa in l1: df_[f'{taxa}_l1'] = df_[taxa].shift(1)
  lx = list(diag[(diag['p-valor'] <= 0.05) & (diag['relevancia_IV']<=10)]['taxas'])
  df_.dropna(inplace=True)
  df_.drop(columns=l1+lx, inplace=True)
  return df_

# redefinição da base com tratamento de endogeneidade
df_glm = trat_endog(diag_endog(df_glm, df_glm.columns[:-2]), df_glm)

# equação linear do modelo
def formula(df_):
  c = df_.columns.drop(['Sinistralidade', 'logPremioEmitido'])
  return f'Sinistralidade ~ {'+'.join(c for c in c)}'

# Modelo GLM Poisson
glm_poisson = smf.glm(formula=formula(df_glm), data=df_glm,
```

```
                      family=sm.families.Poisson(),
                      offset=df_glm['logPremioEmitido']
                      )

res_glm_poisson = glm_poisson.fit()
print(res_glm_poisson.summary().as_text())

# teste de heterocedastidade e autocorrelacao dos residuos
import statsmodels.api as sm

def diag_het_acorr(modelo):

  nome = modelo.model.family.__class__.__name__
  print(f'Modelo GLm - {nome}')

  # Breusch Pagan
  print('\nBreusch Pagan (p-valor):')
  print(sm.stats.diagnostic.het_breuschpagan(
      modelo.resid_pearson,
      sm.add_constant(df_glm.drop(columns=['Sinistralidade', 'logPremioEmitido'])))[1])

  # White
  # sm.stats.diagnostic.het_white(
  #     modelo.resid_pearson,
  #     sm.add_constant(df_glm.drop(columns=['Sinistralidade'])))

  # Ljung-Box
  print('\nLjung-Box (p-valor):')
  print(sm.stats.diagnostic.acorr_ljungbox(
      modelo.resid_pearson,
      lags=[12]).lb_pvalue[12])
```

```
# diagnóstico - GLM - Poisson
diag_het_acorr(res_glm_poisson)

# comparação entre valores observados e ajustados no modelo GLM
def graf_comp(modelo):
  plt.subplots(figsize=(5.5, 3.0))
  nome = modelo.model.family.__class__.__name__
  df_glm['Sinistralidade'].plot(label='Sinistralidade observada', lw=1, style='k--')
  modelo.fittedvalues.plot(label = f'Sinistralidade GLM - {nome}')
  modelo.resid_pearson.plot(label='residuos Pearson')
  plt.xlabel('')
  plt.legend()
  plt.tight_layout()
  plt.show()

  #plt.subplots(figsize=(5.5, 3.0))
  #plt.scatter(res_glm_poisson.fittedvalues, res_glm_poisson.resid_pearson)

# grafico comparativo GLM-Poisson
graf_comp(res_glm_poisson)

# Modelo Binomial Negativa
glm_negbinom = smf.glm(formula=formula(df_glm), data=df_glm,
                       family=sm.families.NegativeBinomial(alpha=1.0),
                       offset=df_glm['logPremioEmitido']
                       )

res_glm_negbinom = glm_negbinom.fit()
print(res_glm_negbinom.summary().as_text())
```

```
# diagnóstico residuos GLM - Binominal Negativa
diag_het_acorr(res_glm_negbinom)

# grafico comparativo GLM - Binomial Negativa
graf_comp(res_glm_negbinom)

# Modelo Gamma
glm_gamma = smf.glm(formula=formula(df_glm), data=df_glm,
                    family=sm.families.Gamma(link=sm.families.links.Log()),
                    offset=df_glm['logPremioEmitido']
                    )

res_glm_gamma = glm_gamma.fit()
print(res_glm_gamma.summary().as_text())

# diagnóstico residuos GLM - Gamma
diag_het_acorr(res_glm_gamma)

# grafico comparativo GLM - Gamma
graf_comp(res_glm_gamma)

# Modelo Inverse Gaussian
glm_invgauss = smf.glm(formula=formula(df_glm), data=df_glm,
                       family=sm.families.InverseGaussian(link=sm.families.links.Log()),
                       offset=df_glm['logPremioEmitido']
                       )

res_glm_invgauss = glm_invgauss.fit()
print(res_glm_invgauss.summary().as_text())

# diagnóstico residuos GLM - Gaussiana Inversa
```

```
diag_het_acorr(res_glm_invgauss)

# grafico comparativo GLM - Gaussiana Inversa
graf_comp(res_glm_invgauss)

# Modelo Tweedie
glm_tweedie = smf.glm(formula=formula(df_glm), data=df_glm,
                      family=sm.families.Tweedie(link=sm.families.links.Log(),
                                                 var_power=1.5),
                      offset=df_glm['logPremioEmitido']
                      )

res_glm_tweedie = glm_tweedie.fit()
print(res_glm_tweedie.summary().as_text())

# diagnóstico residuos GLM - Tweedie
diag_het_acorr(res_glm_tweedie)

# grafico comparativo GLM - Tweedie
graf_comp(res_glm_tweedie)

# comparaçao entre modelos GLM
modelos = [res_glm_poisson, res_glm_negbinom, res_glm_gamma,
           res_glm_invgauss, res_glm_tweedie]

for modelo in modelos:
  print(f'AIC Modelo - {modelo.model.family.__class__.__name__}: {modelo.aic}')

# MODELO CLUSTERING

# biblioteca
```

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score

# base do modelo
df_cluster = (dados.groupby('coenti').apply(lambda x: x[taxas].iloc[12:].mean()))
print(df_cluster.describe().T.round(2).to_string())

# transformação de valores em escala min-max
df_cluster_mm = MinMaxScaler().fit_transform(df_cluster)
df_cluster_mm = pd.DataFrame(df_cluster_mm, columns=df_cluster.columns)
df_cluster_mm.index = df_cluster.index

# escolha do número ótimo de agrupamentos
fig, (eixo1, eixo2) = plt.subplots(1, 2, figsize=(6.5, 2.5))
wcss = []
silhueta = []
for n in range(2, 20):
  km = KMeans(n_clusters=n, random_state=42).fit(df_cluster_mm)
  wcss.append(km.inertia_)
  silhueta.append(silhouette_score(df_cluster_mm, km.labels_))

eixo1.plot(range(2, 20), wcss, marker='o', linestyle='--')
eixo1.set_title('Método do Cotovelo')
eixo2.plot(range(2, 20), silhueta, marker='o', linestyle='--')
eixo2.set_title('Método da Silhueta')
plt.tight_layout()
plt.show()

# modelo KMEANS com 7 grupos
```

```
km = KMeans(n_clusters=7, random_state=42)
km.fit(df_cluster_mm)

# analise grafica de dispersão dos grupos
df_cluster_pca = pd.DataFrame(PCA(2, random_state=42).fit_transform(df_cluster))
df_cluster_pca.columns = ['PCA1', 'PCA2']

fig, eixo = plt.subplots(figsize=(7.5, 3.5))
eixo.scatter(data=df_cluster_pca, x='PCA1', y='PCA2', c=km.labels_, cmap='viridis')
for i, coenti in enumerate(list(df_cluster.index)):
  eixo.text(df_cluster_pca['PCA1'][i], df_cluster_pca['PCA2'][i],
            str(coenti), fontsize=6, ha='right')
eixo.set_xlabel('PCA1')
eixo.set_ylabel('PCA2')
plt.tight_layout()
plt.show()

# incluir informação de agrupamento na base de dados
df_grupos = pd.DataFrame({'coenti': df_cluster.index, 'grupo': km.labels_})
dados = pd.merge(dados, df_grupos, on='coenti', how='left')

# dinamicas de sinistralidade por agrupamento
fig, eixo = plt.subplots(4, 2, figsize=(6.5, 6.5))
fig.delaxes(eixo[3, 1])
for i in range(7):
  df = dados[dados['grupo'] == i].groupby('damesano')[taxas].mean().iloc[12:]
  df = df.asfreq('MS')
  df.Sinistralidade.plot(ax=eixo[i//2, i%2])
  eixo[i//2, i%2].set_title(f'Sinistralidade - Grupo {i}')
  eixo[i//2, i%2].set_xlabel('')
  eixo[i//2, i%2].set_ylabel('')
```

```
  eixo[i//2, i%2].spines[['top', 'right']].set_visible(False)
  plt.setp(eixo[i//2, i%2].get_xticklabels(), rotation=45, ha='right')

plt.tight_layout()
plt.show()

# dicionário de bases para modelagem GLM-Tweedie com agrupamento
df_glm_cluster = {}

for i in range(7):
  df_filtro = dados[dados['grupo'] == i]
  df_filtro_media = df_filtro.groupby('damesano')[taxas].mean().iloc[12:]
  df_filtro_media['logPremioEmitido'] = np.log(
      df_filtro.groupby('damesano')['premio_de_seguros_12m_inf'].mean().iloc[12:])
  df_glm_cluster[i] = df_filtro_media

# inclusão de componente autorregressivo da sinistralidade nos grupos
for i in range(7):
  autorr_sinist(df_glm_cluster[i])

# redefinição das variaveis nos grupos com tratamento de endogeneidade
for i in range(7):
  df_glm_cluster[i] = trat_endog(diag_endog(df_glm_cluster[i], df_glm_cluster[i].columns[:-2]),
                                 df_glm_cluster[i])

# Modelos Tweedie com agrupamento
res_glm_tweedie_cluster = []
for i, df in df_glm_cluster.items():
  nneg = df[df['Sinistralidade'] <= 0].shape[0]
  if nneg > 0:
    # print(f'Cluster {i}: {nneg} valores não positivos para Sinistralidade')
```

```
    df.loc[df['Sinistralidade'] <= 0, 'Sinistralidade'] = 1e-6

  glm_tweedie_cluster = smf.glm(formula=formula(df), data=df,
                                family=sm.families.Tweedie(link=sm.families.links.Log(),
                                                           var_power=1.5),
                                offset=df['logPremioEmitido']
                                )

  res_glm_tweedie_cluster.append(glm_tweedie_cluster.fit(cov_type='HC3'))
  # print(f"\nModelo GLM Cluster: {i}")
  # print(glm_tweedie_cluster.fit(cov_type='HC3').summary().as_text())

# tabelas de termos e coeficientes
tb = [res_glm_tweedie_cluster[i].summary2().tables[1] for i in range(7)]
for i in range(7):
  tb[i]['Grupo'] = i
  tb[i]['Lag'] = np.zeros(tb[i].shape[0])
  tb[i].loc[[j for j in tb[i].index if '_l1' in j], 'Lag'] = 1

# coeficientes significativos
coef_signif = pd.concat(tb)
coef_signif.loc[coef_signif['P>|z|'] > 0.05,'Coef.'] = 0
coef_signif.loc[coef_signif['Coef.'] == 0, 'Lag'] = 0
coef_signif.index = coef_signif.index.str.replace('_l1', '')
coef_signif = coef_signif[['Coef.', 'Grupo', 'Lag']].reset_index()
coef_signif.columns = ['Variavel', 'Coeficiente', 'Agrupamento', 'Lag']

# ordem de importancia entre os fatores
zeros = [coef_signif[(coef_signif['Variavel']==c)&(coef_signif['Coeficiente']==0)].shape[0]
         for c in coef_signif.Variavel.unique()]
ordem = pd.Series(dict(zip(coef_signif.Variavel.unique(), zeros))).to_frame().sort_values(0).index
```

```
# gráfico de coeficientes significativo por grupo
fig, eixo = plt.subplots(3, 3, figsize=(7.5, 4.5))
for i, c in enumerate(ordem):
  df = coef_signif[coef_signif['Variavel'] == c]
  ax = eixo[i//3, i%3]
  g = sns.barplot(data=df, x='Agrupamento', y='Coeficiente', ax=ax, palette='viridis')
  ax.set_title(c)
  ax.set_xlabel('Grupo')
  ax.set_ylabel('Coeficiente')
  ax.axhline(0, color='gray', linestyle='--', linewidth=0.8)

  # anotação L1 para os coeficientes de regressor em defasagem
  for index, row in df.iterrows():
    if row['Lag'] == 1:
      y_pos = ax.get_ylim()[1] * 0.85 if row['Coeficiente'] > 0 else ax.get_ylim()[0] * 0.50
      ax.text(row['Agrupamento'], y_pos, 'L1', color='black', ha='center', va='top', fontsize=8)

plt.tight_layout()
plt.show()

# calculo do efeito multiplicador em variação percentual
coef_efeito = coef_signif.copy()
coef_efeito.loc[coef_efeito['Coeficiente'] != 0, 'Coeficiente'] = (
    np.exp(coef_signif.loc[coef_signif['Coeficiente'] != 0, 'Coeficiente'] * 0.01) - 1)

# gráfico de efeitos multiplicadores para coeficientes significativos
import matplotlib.ticker as mticker

fig, eixo = plt.subplots(3, 3, figsize=(7.5, 4.5))
for i, c in enumerate(ordem):
```

```
  df = coef_efeito[coef_efeito['Variavel'] == c]
  ax = eixo[i//3, i%3]
  g = sns.barplot(data=df, x='Agrupamento', y='Coeficiente', ax=ax, palette='viridis')
  ax.set_title(c)
  ax.set_xlabel('Grupo')
  ax.set_ylabel('% Sinistralidade')
  ax.axhline(0, color='gray', linestyle='--', linewidth=0.8)
  ax.yaxis.set_major_formatter(mticker.PercentFormatter(xmax=1.0, decimals=0))

  # anotação L1 para os coeficientes de regressor em defasagem
  for index, row in df.iterrows():
    if row['Lag'] == 1:
      y_pos = ax.get_ylim()[1] * 0.85 if row['Coeficiente'] > 0 else ax.get_ylim()[0] * 0.50
      ax.text(row['Agrupamento'], y_pos, 'L1', color='black', ha='center', va='top', fontsize=8)

plt.tight_layout()
plt.show()
```