PUC Minas
E&G Economia & Gestão
ISSN 1984-6606

MODELOS DE POISSON INFLADA DE ZEROS E BINOMIAL NEGATIVA INFLADA DE ZEROS NA PREVISÃO DE SINISTRO DE AUTOMÓVEIS

ZERO-INFLATED POISSON AND ZERO-INFLATED NEGATIVE BINOMIAL MODELS TO PREDICT AUTOMOBILE CLAIMS

Natália Zaniboni
Universidade de São Paulo – USP
nzaniboni@usp.br

Alessandra Montini
Universidade de São Paulo – USP
amontini@usp.com.br

Submissão: 21/10/2014
Aprovação: 25/09/2015

RESUMO

As seguradoras estão constantemente aprimorando seus modelos de sinistralidade para obter preços mais competitivos. A distribuição de probabilidade dos sinistros possui alta frequência no valor zero, possibilitando o ajuste das distribuições Poisson Inflada de Zeros (ZIP) e Binomial Negativa Inflada de Zeros (ZINB). O objetivo do trabalho é preencher duas lacunas da literatura em relação a estes modelos: verificar se fatores humanos e externos afetam o número de sinistros e estimar tanto a probabilidade de ocorrência do sinistro quanto o número de sinistros por meio do ajuste de modelos ZIP e ZINB para o número de sinistros das carteiras de seguro de automóveis. A base de dados foi extraída do sistema AUTOSEG, da SUSEP, e as variáveis sexo, estado de residência e idade do condutor, categoria tarifária e ano do modelo do automóvel, número de expostos e importância segurada média foram significativas nos modelos ajustados a um nível de 10%. O modelo de ZIP apresentou erro quadrático médio menor que os modelos ZINB, Poisson e Binomial Negativa.

Palavras-chave: Poisson Inflada de Zeros, Binomial Negativa Inflada de Zeros, sinistros de automóveis.

ABSTRACT

Insurance companies are constantly improving their claims predicting models for more competitive prices. The probability distribution has a high frequency at the value zero, allowing the adjustment of Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) distributions. The main objective of this paper is to fill two gaps in the literature: analyze if human and external factors affect the number of claims and estimate both the probability of occurrence and the number of claims in a portfolio by fitting ZIP and ZINB models in an automobile insurance portfolio. The variables gender, state of residence and age of the driver, tariff category and year of the car, number of exposed and insured amount average were significant at a level of 10%. The ZIP model showed lower mean square error compared to ZINB, Poisson and Negative Binomial models.

Key-words: Zero-Inflated Poisson, Zero-Inflated Negative Binomial, automobile claims.

1. Introdução

O mercado segurador de automóveis é muito competitivo e está em pleno crescimento. O percentual de automóveis do país que possuem seguro de auto não chega a 40%. A precificação é muito importante na competitividade de uma empresa seguradora, pois, além da massificação da carteira, é necessário um excelente modelo de precificação para a obtenção de preços competitivos sem subestimar as possíveis perdas da seguradora com os sinistros. O ajuste de uma adequada distribuição de probabilidade dos sinistros é fundamental para que o preço praticado seja o menor possível, fazendo com que a seguradora possa captar um número cada vez maior de segurados.

Com relação à distribuição dos sinistros tem-se que existe alta concentração de segurados que não sinistraram no período de vigência da sua apólice. Por este motivo a distribuição de probabilidade possui alta frequência no valor zero sinistro. Desta forma, esta distribuição é inflada no valor zero, possibilitando o ajuste das distribuições Poisson Inflada de Zeros (ZIP) (LAMBERT, 1992) e Binomial Negativa Inflada de Zeros (ZINB) (LAWLESS, 1987). Técnicas de previsão da contagem de sinistros tem sido o tema de muitos trabalhos na literatura atuarial, porém poucos trabalhos tratam esta característica da sua distribuição (BOUCHER *et al.*, 2009).

As distribuições Infladas de Zeros são muito utilizadas em dados de contagem com alta concentração no valor zero e a literatura contém diversos exemplos. Podem-se destacar os dados de defeitos em equipamentos de fabricação (LAMBERT, 1992), acidentes de caminhão (MIAOU, 1994), número de visitas a médicos e hospitais (GURMU, 1997), número de dentes deteriorados ou faltantes em pacientes (BÖHNING *et al.*, 1995), número de patentes de produtos estudados pela área de pesquisa e desenvolvimento (P&D) de uma empresa (CREPON; DUGUET, 1997), animais em espécies raras (WELSH *et al.*, 1996), número de viagens de navio para entretenimento (GURMU; TRIVEDI, 1996), comportamento de risco em homossexuais (HEILBRON, 1989), contagem de organismos raros em ecologia (MARTIN *et al.*, 2005) e acidentes de veículos (LORD *et al.*, 2005).

A respeito de modelos de sinistros de automóveis, Martins (2012) indica que fatores humanos e externos afetam a probabilidade de ocorrência de acidentes de automóveis, e ambas as classes de variáveis devem ser consideradas. Bortoluzzo *et al.* (2011) apontam a necessidade de se estimar tanto a probabilidade de ocorrência do sinistro quanto o número de sinistros de uma carteira.

Este trabalho irá ajustar as distribuições Poisson Inflada de Zeros e Binomial Negativa Inflada de Zeros para o número de sinistros das carteiras de seguro de automóveis e serão comparadas as previsões geradas para identificar o modelo melhor ajustado. As contribuições do trabalho para a literatura são: (1) propor um ajuste de uma distribuição mais adequada (Poisson Inflada de Zeros e Binomial Negativa Inflada de Zeros) à quantidade de sinistro de automóveis, (2) estimar tanto a probabilidade de ocorrência do sinistro quanto o número de sinistros por meio das distribuições propostas (BORTOLUZZO *et al.*, 2011) e (3) avaliar o impacto de fatores humanos (características dos condutores de veículos) e externos (características do ambiente) na quantidade de sinistros de automóveis (MARTINS, 2012).

Notou-se que as variáveis estado, idade e sexo do condutor, categoria tarifária e ano do modelo do veículo (características dos condutores e dos veículos), quantidade de veículos expostos e importância segurada média da carteira (características do ambiente) foram significativas para a previsão do número de sinistros da carteira. Os modelos que consideram as distribuições inflada de zeros(ZIP e ZINB) apresentaram menor erro quadrático médio na previsão comparado aos modelos Poisson e Binomial Negativa, sendo mais indicados para projetar o número de sinistros de uma carteira de automóveis.

2. Referencial Teórico

2.1. Modelos Aplicados a Dados de Sinistros

Os estudos sobre seguro de automóveis utilizando modelagem estatística se iniciaram na década de 60 (BAILEY; SIMON, 1960; JUNG, 1968) e 70 (BENNETT, 1978; JOHNSON; HEY, 1971) e abordavam, principalmente, a análise do retorno financeiro de uma carteira de segurados.

Junior (2003) construiu um modelo de Poisson para previsão da taxa de ocorrência de sinistros utilizando o período de subscrição e período de ocorrência como variáveis explicativas, indicando que ambas são significativas a um nível de 5%.

Filho e Lugon (2004) utilizaram análise discriminante para determinar quais fatores afetam a ocorrência de sinistro de automóvel do segurado. A base de dados utilizada continha 21.024 registros, e o estudo foi realizado no período de janeiro a dezembro de 2000. Os fatores faixa etária, estado civil, ocupação, nível de escolaridade e registro no Serasa afetaram

a ocorrência do sinistro do segurado. O modelo ajustado apresentou um baixo percentual de assertividade, indicando a necessidade de inclusão de novas variáveis ao modelo.

Viaene et al. (2005) utilizaram um modelo de redes neurais Bayesianas para prever a quantidade de fraudes em sinistros de seguros de automóveis com proteção a danos pessoais em dados de uma seguradora em Massachusetts (EUA) no ano de 1993.

Delgado (2011) ajustou uma distribuição Binomial Negativa ao número de sinistros de veículos de uma Companhia Seguradora em Portugal, utilizando uma carteira com dados de dezembro de 2009. A distribuição Binomial Negativa se ajustou de forma mais adequada do que a distribuição de Poisson.

Bortoluzzo et al. (2011) ajustou modelos da família exponencial e modelos Gaussianos inversos zero-ajustados à distribuição de número de sinistros de automóveis de uma Companhia seguradora. Idade, fabricante, país de fabricação e tipo do veículo afetam a probabilidade e o número de ocorrências de sinistros. As duas distribuições foram bem ajustadas aos dados.

Martins (2012) utilizou dados de sinistros de automóveis do grupo Caixa Seguros em Portugal do período de 2000 a 2010. Aproximadamente 80% das apólices não apresentaram sinistros, porém não foram consideradas as distribuições infladas na modelagem dos dados. Idade, sexo, tipo do veículo, idade do veículo, região que reside o condutor e número de veículos em trânsito influenciam a previsão do número de sinistros para todos os modelos ajustados (modelo logístico, distribuição de Poisson e Binomial Negativa). O modelo de Poisson apresentou pior desempenho entre os modelos ajustados.

2.2. Distribuições Infladas de Zero

Considerando o número de sinistros de automóveis no período em estudo (Y) um processo de contagem com grande frequência de zeros. A probabilidade de haver zero sinistro é p e de haver algum sinistro é 1- p (HEILBRON, 1994), conforme a expressão (1).

$$\begin{aligned} P(Y = 0) &= p \\ P(Y \neq 0) &= 1 - p \end{aligned} \qquad (1)$$

Heilbron (1989) propôs, para dados com excesso de zeros, ajustar uma distribuição de probabilidade em duas fases. Na primeira fase estima-se a probabilidade do valor ser diferente

de zero (dada pela expressão (1)). Na segunda fase estima-se o valor da variável aleatória considerando uma distribuição da família exponencial.

Os modelos são ajustados utilizando modelos lineares generalizados e assumem que as distribuições das duas fases são independentes. Heilbron (1989) aplicou a metodologia a dados de comportamentos de alto risco em homossexuais.

2.2.1. Poisson Inflada de Zeros (ZIP)

Cohen (1963), Singhn (1963) e Johnson e Koltz (1969) apresentaram a distribuição Poisson Inflada de Zeros (ZIP), que se aplica a dados de contagem quando há uma frequência excessiva do valor zero. Esta distribuição é baseada em uma mistura de duas distribuições, assim como Heilbron (1989), onde uma componente representa a probabilidade dos valores iguais zero e outra componente representa a distribuição de Poisson, ajustada aos valores maiores do que zero. Os parâmetros da distribuição são definidos pela proporção de zeros (p) e pela média da distribuição Poisson (λ) excluindo os zeros. A distribuição de probabilidade é dada pela expressão (2).

$$P(Y=y)=\begin{cases} p+(1-p)e^{-\lambda} & , y=0 \\ (1-p)\dfrac{e^{-\lambda}\lambda^{y}}{y!} & , y>0 \end{cases} \qquad (2)$$

em que y é o número de sinistros na carteira ($y \geq 0$), p é a proporção de zeros e λ é o parâmetro da média.

Lambert (1992) propôs o modelo de Poisson Inflada de Zeros utilizando covariáveis. Os parâmetros p e λ são estimados via modelos lineares generalizados. O autor comparou os modelos de Poisson Inflada de Zeros e Binomial Negativa para uma base de defeitos em equipamentos. Notou-se que a Poisson Inflada de Zeros obteve melhores previsões. A Binomial Negativa superestimou os valores preditos.

Miaou (1994) utilizou modelos de Poisson, Poisson Inflada de Zeros e Binomial Negativa aplicados a dados de acidentes com caminhões em rodovias. Indicou, através da qualidade do ajuste, que o modelo de Poisson Inflada de Zeros é bem ajustado quando há excesso de dados com contagem igual a zero. Segundo o autor, os modelos de Binomial Negativa devem ser usados com cautela. Se a superdispersão dos dados é moderada ou alta, tanto o modelo de

Binomial Negativa quanto a regressão Poisson Inflada de Zeros podem ser utilizados, mas o modelo Poisson Inflada de Zeros é mais indicado quando há excesso de zeros.

Böhning et al. (1995) aplicaram o modelo de Poisson Inflada de Zeros em dados de epidemologia dental de crianças de uma área urbana de Belo Horizonte, onde foram medidos o número de dentes faltantes, em quedas ou preenchidos. Vários zeros foram encontrados na base, que representam as crianças sem problemas. O ajuste de uma distribuição de Poisson não foi adequado neste caso pois a média e a variância são muito diferentes. O autor incluiu no modelo as covariáveis método de prevenção, sexo e cor da pele, e encontrou um ganho do modelo de Poisson Inflado de Zeros de dois componentes em comparação aos modelos Log-Linear e Log-Linear Inflado de Zeros.

2.2.2. Binomial Negativa Inflada de Zeros (ZINB)

A distribuição Binomial Negativa Inflada de Zeros (ZINB), proposta por (LAWLESS, 1987), tem a mesma característica da distribuição de Poisson Inflada de Zeros, em que existe uma probabilidade, p, de a observação assumir o valor zero e a probabilidade, 1-p, de a observação assumir um valor diferente de zero. A probabilidade da observação assumir um valor diferente de zero é modelada pela distribuição Gama (KHAMKONG, 2010). A distribuição Binomial Negativa Inflada de zeros pode ser descrita pela expressão (3).

$$P(Y=y)=\begin{cases} p+(1-p)\left(\dfrac{1}{1+\sigma\lambda}\right)^{1/\sigma} & , y=0 \\ (1-p)\dfrac{\Gamma\left(y+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(1+y)}\left(\dfrac{\sigma\lambda}{1+\sigma\lambda}\right)^{y}\left(\dfrac{1}{1+\sigma\lambda}\right)^{1/\sigma} & , y>0 \end{cases} \quad (3)$$

em que y é o número de sinistros na carteira (y≥0), p é a proporção de zeros, λ é o parâmetro da média, σ é o parâmetro de dispersão e Γ(.) é a função Gama.

O modelo ZINB com covariáveis foi apresentado por Wang (2003), que aplicou o modelo na previsão do número de consultas com médicos na Austrália. Yau et al. (2003) utilizaram a distribuição ZINB em dados de internação hospitalar devido a problemas no pâncreas na Austrália. Indicaram que idade e o número de diagnósticos afetam a probabilidade zero inflada. No componente Binomial Negativa, o estado civil, tipo de admissão, categoria do tratamento e número de diagnósticos afetam no tempo de internação do paciente. Segundo Yau et al. (2003), o modelo ZINB se ajustou bem aos dados.

Santos et al. (2011) aplicaram o modelo ZINB em dados de duração do absenteísmo-doença segundo Licenças para Tratamento de Saúde (LTS), da Secretaria de Estado da Saúde de Santa Catarina. Os autores apontaram que este modelo se ajustou bem aos dados, apresentando menores discrepâncias (AIC e BIC) comparado aos modelos tradicionais para dados de contagem.

A Tabela 1 apresenta a média e variância das distribuições Poisson, Poisson Inflada de Zeros, Binomial Negativa e Binomial Negativa Inflada de Zeros. Nota-se que a distribuição Binomial Negativa Inflada de Zeros acomoda a sobredispersão em sua variância (parâmetro σ), com a mesma média da Poisson Inflada de Zeros.

Tabela 1: Média e variância das distribuições

Distribuição	Média	Variância
Poisson	λ	λ
Poisson Inflada de Zeros	$(1-p)\lambda$	$(1-p)\lambda(1+p\lambda)$
Binomial Negativa	λ	$\lambda + \lambda\sigma^2$
Binomial Negativa Inflada de Zeros	$(1-p)\lambda$	$(1-p)\lambda(1+(\sigma+p)\lambda)$

Fonte: elaborado pelo autor

2.2.3. Comparativos entre ZIP e ZINB

Fernandez e Botelho (2005) utilizaram os modelos Poisson, Binomial Negativa, ZIP e ZINB, com e sem variáveis explicativas, em dados de consumo de cigarros. O modelo ZINB melhor se ajustou aos dados.

Yip e Yau (2005) aplicaram as distribuições ZIP, ZINB, ZIGP e ZIDP em dados de sinistros de automóveis (retirados do banco de dados do SAS Enterprise Miner) e o modelo ZIP proporcionou um bom ajuste aos dados. O objetivo de uso do carro, estado civil, local de residência, renda e sexo dos condutores se mostraram significativos nos modelos ajustados.

Carvalho e Lavor (2008) analisaram o impacto de características socioeconômicas do indivíduo e da região no processo da vitimização criminal repetida. Segundo os autores o modelo ZINB apresentou melhores resultados, medidos através da máxima verossimilhança, comparado aos modelos Poisson, Poisson Inflado de Zeros e Binomial Negativa.

Boucher et al. (2009) propôs um modelo combinado ZIP-Gamma para prever o número de sinistros de automóveis de uma seguradora na Espanha. As variáveis explicativas utilizadas foram sexo do condutor, tempo em que o cliente está na empresa, idade do condutor e a potência do automóvel. O modelo ZIP-Gamma se ajustou melhor aos dados que as distribuições Binomial Negativa e ZINB.

ISSN 1984-6606

Souza (2011) indicou que, aplicado em dados de prevalência de estresse pós-traumático em militares, o modelo Binomial Negativa Inflada de Zeros é mais apropriado que os modelos Poisson, Poisson Inflado de Zeros e Binomial Negativa. Utilizou a mensuração do observado e predito para comparar os modelos.

3. Base de Dados

A base de dados foi extraída do sistema AUTOSEG, da SUSEP (Superintendência de Seguros Privados), no dia 4/01/2012. Este sistema é gratuito e permite realizar consultas sobre dados de seguros de automóveis referentes a apólices vigentes no período de análise. O banco de dados ainda disponibiliza as informações desejadas classificadas por perfil do segurado, do automóvel e da região de risco. As sociedades seguradoras realizam o envio destas informações semestralmente à SUSEP. Utilizou-se o segundo semestre de 2010, informação mais recente disponível na data da extração.

A variável resposta da base de dados é o número de sinistros causados por roubo ou furto, colisão parcial, colisão perda total, incêndio e outros motivos. As variáveis explicativas contidas na base de dados e a sua relação esperada com a variável resposta são apresentadas na Tabela 2.

Tabela 2: Variáveis explicativas contidas na base de dados e relação esperada com o número de sinistros

Variável	Relação esperada com o número de sinistros
Ano do modelo do carro	Quanto mais antigo o carro, maior a chance do sinistro (MARTINS, 2012).
Modelo do carro (incluindo número de portas, motor, potência, etc.)	A chance do sinistro diminui em carros de luxo (BORTOLUZZO et al., 2011).
Categoria tarifária do automóvel, classificada como: 1. Passeio nacional 2. Passeio importado 3. Pick-up (nacional e importado) 4.Veículo de Carga (nacional e importado) 5. Motocicleta (nacional e importado) 6. Ônibus (nacional e importado) 7. Utilitários (nacional e importado) 8. Outros	Categorias de passeio tem maior chance de sinistros que demais (MARTINS, 2012).
Número de contratos de seguro expostos na carteira	Quanto maior o número de exposições maior o número de sinistros (MARTINS, 2012)
Faixa etária do principal condutor do veículo, classificada em: 1. Entre 18 e 25 anos 2. Entre 26 e 35 anos 3. Entre 36 e 45 anos 4. Entre 46 e 55 anos 5. Maior que 55 anos	Pessoas mais novas tendem a ter uma chance maior de sinistro (BOUCHER et al., 2009)
Importância segurada média dos contratos expostos	Veículos mais caros tendem a ter menor chance de sinistro, dado que a chance do sinistro diminui em carros de luxo (BORTOLUZZO et al., 2011)
Soma dos prêmios pagos dos contratos expostos	Veículos mais caros, que pagam maior prêmio, tendem a ter maior chance de sinistro (BORTOLUZZO et al., 2011)
Região do risco	Regiões metropolitanas tendem a ter maior chance de sinistros (MARTINS, 2012)
Sexo do principal condutor, classificado em: 1. Masculino 2. Feminino 3. Jurídica 4. Sem Informação	Homens tendem a ter maior chance de sinistros (BOUCHER et al., 2009; MARTINS, 2012)

Uma linha da base de dados representa a combinação das variáveis categóricas, ou seja, indica a carteira do seguro de automóvel. As variáveis de sinistros, exposições, prêmios, importância segurada e indenizações representam a soma ou a média destes dados na combinação.

Foram agrupados os códigos do modelo do carro por seu modelo principal, ou seja, todas as versões de um mesmo modelo de automóvel foram agrupadas. A base após o agrupamento continha 751.631 registros. Foram filtradas somente as combinações que possuem contratos ativos, ou seja, expostos (número de contratos) > 0, totalizando uma base de 726.156 registros (carteiras).

Os modelos a serem estimados, que preveem a probabilidade de ocorrência do sinistro e a quantidade de sinistros, respectivamente, estão descritos nas expressões (4) e (5):

$\text{Probabilidade de não ocorrência do sinistro} = \text{p} = \frac{1}{1 + \text{e}^{\beta_0 + \beta_1 \text{Ano do modelo} + \beta_2 \text{Modelo} + \cdots}}$	(4)
$\text{Quantidade de Sinistros} = (1 - \text{p}) * \text{e}^{\gamma_0 + \gamma_1 \text{Ano do modelo} + \gamma_2 \text{Modelo} + \cdots}$	(5)

em que β_k é o coeficiente estimado no modelo da expressão (4) para a variável explicativa k e γ_k é o coeficiente estimado no modelo da expressão (5) para a variável explicativa k.

4. Resultados

4.1. Análise Descritiva

A partir da base inicial, de 726.156 registros, foram excluídas informações sem preenchimento para alguma variável e formou-se a nova base de dados denominada base filtrada. A variável região foi agrupada por estado.

A partir da base filtrada, com 609.885 observações, foram excluídas observações com número de sinistros maior ou igual a 7 (percentil 95%) para evitar distorções na análise. A tabela final resultante apresentou 578.164 observações. Algumas estatísticas descritivas das variáveis explicativas após todos os filtros e análises da variável de interesse deste estudo serão apresentadas nas tabelas seguintes.

A Tabela 3 apresenta a distribuição de frequências das variáveis: Década do Modelo, Categoria Tarifária, Faixa de Idade, Estado e Sexo. Nota-se que a maioria dos veículos pertence à categoria Nacional, com modelos entre os anos 2006 e 2009. Os condutores se encontram, em sua maioria, no estado de São Paulo e são predominantemente do sexo masculino.

Tabela 3: Distribuição de Frequência

Variável	Categoria	Frequência	%
Década do Modelo	< 1950	11	0%
	1951 a 1960	67	0%
	1961 a 1970	261	0%
	1971 a 1980	2.005	0%
	1981 a 1990	12.787	2%
	1991 a 2000	103.295	18%
	2001 a 2005	175.878	30%
	2006 a 2009	214.498	37%
	2010 a 2011	69.362	12%
Categoria Tarifária	Passeio nacional	258.996	45%
	Passeio importado	119.963	21%
	Pick-up (nacional e importado)	130.308	23%
	Veículo de Carga (nacional e importado)	31.462	5%
	Motocicleta (nacional e importado)	23.143	4%
	Ônibus (nacional e importado)	1.464	0%
	Utilitários (nacional e importado)	1.128	0%
	Outros	11.700	2%
Faixa de Idade	Entre 18 e 25 anos	55.260	10%
	Entre 26 e 35 anos	115.564	20%
	Entre 36 e 45 anos	138.196	24%
	Entre 46 e 55 anos	136.503	24%
	Maior que 55 anos	132.641	23%
Região	Sudeste	242.110	42%
	Sul	161.614	28%
	Norte	110.981	19%
	Centro-Oeste	55.462	10%
	Nordeste	7.997	1%
Sexo	Feminino	212.046	37%
	Masculino	277.836	48%
	Jurídico	88.282	15%

Fonte: elaborado pelo autor

A Tabela 4 apresenta o mínimo, percentil 1%, percentil 5%, média, percentil 95%, percentil 99% e máximo das variáveis quantidade de veículos expostos (contratos de segurados), a importância segurada média e o prêmio pago pelos expostos. Nota-se que a distribuição do prêmio possui uma assimetria, pois o máximo é muito superior à mediana.

Tabela 4: Estatísticas descritivas da quantidade de veículos expostos, importância segurada média e prêmio pago pelos expostos

Variável	Mínimo	Perc. 1%	Perc. 5%	Mediana	Média	Perc. 95%	Perc. 99%	Máximo
Expostos	0,00*	0,02	0,10	0,95	3,70	17,65	40,99	280,38
Importância Segurada (R$mil)	0,00	0,00	1.658,96	30.019,00	40.940,71	113.162,00	200.693,85	967.702,17
Prêmio Pago (R$mil)	0,00**	10,62	53,41	1.278,07	4.458,85	19.634,48	44.830,30	560.578,46

* valor igual a 0,0027
** valor igual a 0,0026
Fonte: elaborado pelo autor

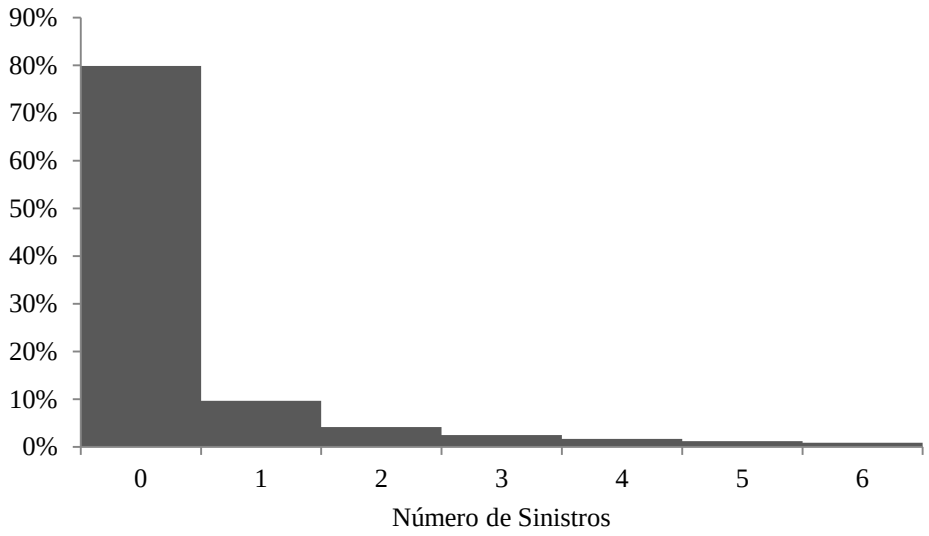
A Tabela 5 apresenta as estatísticas descritivas (mínimo, percentil 1%, percentil 5%, mediana, média, percentil 95%, percentil 99% e máximo) do número de sinistros para a nova tabela, já sem os dados discrepantes. A Figura 1 apresenta o histograma para esta variável. Nota-se que aproximadamente 80% dos dados se concentram no valor zero.

Tabela 5: Estatísticas descritivas do número de sinistros

Variável	Mínimo	Perc. 1%	Perc. 5%	Mediana	Média	Perc. 95%	Perc. 99%	Máximo
Nº Sinistros	0	0	0	0	0,4	3,0	5,0	6,0

Fonte: elaborado pelo autor

Figura 1: Histograma da variável número de sinistros



Fonte: elaborado pelo autor

A Tabela 6 apresenta a distribuição de frequências do número de sinistros. Nota-se que 79,8% dos registros contem zero sinistros e 20,2% apresentou pelo menos um sinistro.

Tabela 6: Distribuição de frequência do número de sinistros

Variável	Total	%
Nº Sinistros = 0	461.658	79,8%
Nº Sinistros > 0	116.506	20,2%

Fonte: elaborado pelo autor

A Tabela 7 apresenta a média e a variância para a distribuição dos dados considerando o número de sinistros maiores do que zero. Nota-se que a média e a variância são muito próximas, caracterizando uma distribuição de Poisson. Logo, tanto os modelos de Poisson Inflada de Zeros quanto o modelo Binomial Negativa Inflada de Zeros podem ser aplicados a este banco de dados.

Tabela 7: Média e variância do número de sinistros maiores que zero

Variável	Média	Variância
Nº Sinistros > 0	2,16	2,15

Fonte: elaborado pelo autor

4.2. Modelos ZIP e ZINB

As variáveis categóricas foram reagrupadas com objetivo de maximizar diferenciação de número de sinistros (zero ou maior que zero), através do critério de Entropia. A Tabela 8 contém as variáveis, as categorias e o indicativo de significância dos modelos obtidos, sendo que o modelo logístico é a primeira fase do modelo (se ajustam as probabilidades) e o modelo discreto é a segunda fase do modelo (se ajusta a distribuição dos dados maiores do que zero).

O método de seleção de variáveis foi o *Backward*, que retira as variáveis não significativas no modelo uma a uma, iniciando com a variável com o maior nível descritivo. Foi considerado um nível de significância de 10%, ou seja, foram mantidas no modelo as variáveis com nível descritivo menor que 10%. A Tabela 8 apresenta os modelos logísticos, que estimam a probabilidade de não ocorrência de sinistro, conforme a expressão (4), para as distribuições ZINB e ZIP. Nestes modelos, quanto maior o coeficiente, maior a probabilidade de não ocorrência de sinistro, ou seja, menor o risco. Os resultados dos modelos logísticos indicam que seguros de automóveis feitos a mulheres e pessoas jurídicas apresentam maior probabilidade de não ocorrência de sinistro, ou menor risco que homens, verificado nos modelos ZIP e ZINB (BOUCHER et al., 2009; MARTINS, 2012). Clientes com 18 a 25 anos apresentam menor probabilidade de não ocorrência de sinistro, ou maior risco que os demais, também verificado em ambos os modelos (BOUCHER et al., 2009). Automóveis da região

Sul apresentam maior probabilidade de não ocorrência de sinistro, ou menor risco, que os demais em ambos os modelos. Segundo Martins (2012), automóveis das regiões mais metropolitanas apresentam maior risco, confirmando os resultados, pois as demais regiões possuem, em sua maior parte, segurados da região Sudeste. Seguros feitos a automóveis da categoria passeio nacional tem maior risco que as demais (MARTINS, 2012). Os automóveis com ano maiores que 2010, ou mais recentes, têm menor risco que os demais, também verificado nos modelos logísticos das distribuições ZIP e ZINB (MARTINS, 2012). Quanto maior a importância segurada média da carteira, maior a probabilidade de não ocorrência de sinistro, pois a chance de sinistro diminui para veículos de luxo, verificado em ambos os modelos (BORTOLUZZO *et al*., 2011).

Tabela 8: Variáveis, categorias, parâmetros e níveis descritivos dos modelos logísticos para ZINB e ZIP

Variável	Categoria	Modelo Logístico	
		Parâmetro ZINB	Parâmetro ZIP
Intercepto		0,4061	0.9597
Sexo	Masculino	-	-
	Feminino e Pessoa Jurídica	0,6608	0,5589
Idade	Entre 18 e 25 anos	-0,2886	-0,3054
	Maior que 25 anos	-	-
Região	Sul	-	-
	Demais regiões	-0,3325	-0,3106
Categoria Tarifária	Passeio nacional	-	-
	Demais	1,0403	0,9315
Ano do Modelo	< 2010	-0,1582	-0,1242
	≥ 2010	-	-
Importância Segurada Média		Não foi significativa	0.0000011

Nota: As categorias que apresentam (-) são as categorias de referência. Desta forma, os demais coeficientes estimados são relacionados à categoria de referência.
Nota 2: O nível de significância considerado foi 10%.
Fonte: elaborado pelo autor

A Tabela 9 apresenta os modelos discretos, que estimam a quantidade de sinistros esperada, conforme a equação (5). Neste modelo, quanto maior o coeficiente, maior o número de sinistros estimado. Os resultados dos modelos indicam que seguros de automóveis feitos a mulheres e pessoas jurídicas apresentam menor quantidade de sinistros estimada, ou menor risco, que homens, verificado nos modelos ZINB e ZIP (BOUCHER et al., 2009; MARTINS, 2012). Clientes com 18 a 25 anos apresentam maior número de sinistros, também verificado em ambos os modelos (BOUCHER et al., 2009). Automóveis da região Sul apresentam menor

quantidade de sinistros estimada para os modelos ZIP e ZINB (MARTINS, 2012). A carteira de automóveis da categoria passeio nacional tem maior número de sinistros esperado para os modelos ZIP e ZINB (MARTINS, 2012). Os automóveis com ano maiores que 2010, ou mais recentes, têm menor risco que os demais, também verificado nos modelos ZIP e ZINB (MARTINS, 2012).

Tabela 9: Variáveis, categorias, parâmetros e níveis descritivos dos modelos discretos para ZINB e ZIP

Variável	Categoria	Modelo Discreto	
		Parâmetro ZINB	Parâmetro ZIP
Intercepto		-0,1066	0,2876
Sexo	Masculino	-	-
	Feminino e Pessoa Jurídica	-0,0352	-0,0275
Idade	Entre 18 e 25 anos	0,3309	0,2839
	Maior que 25 anos	-	-
Região	Sul	-	-
	Demais Regiões	0,1790	0,1533
Categoria Tarifária	Passeio nacional	-	-
	Demais	-0,4120	-0,3544
Ano do Modelo	< 2010	0,0609	0,0504
	≥ 2010	-	-

Nota: As categorias que apresentam (-) são as categorias de referência. Desta forma, os demais coeficientes estimados são relacionados à categoria de referência.
Nota 2: O nível de significância considerado foi 10%.
Fonte: elaborado pelo autor

Os modelos criados foram comparados com os modelos clássicos para previsão de dados de contagem (poisson e binomial negativa) por meio do erro quadrático médio, com o objetivo de verificar se as distribuições infladas de zeros se adequam melhor aos dados de previsão da quantidade de sinistros que as distribuições de contagem sem este ajuste. Caso esta hipótese se confirme, há provavelmente uma super dispersão dos dados da quantidade de sinistros e deve-se utilizar os modelos inflados de zeros ao invés dos modelos de contagem tradicionais (SANTOS et al., 2011).

A Tabela 11 apresenta a estimação destes dois modelos com base nas mesmas variáveis utilizadas para construção dos modelos ZIP e ZINB. Nota-se que as relações das variáveis são as mesmas para os modelos anteriores, em que carteiras de seguro de automóveis pertencentes a homens, entre 18 e 25 anos, de regiões que não a região Sul, da categoria passeio nacional e veículos antigos (anteriores a 2010) possuem maior quantidade de sinistros estimada. Além

disso, quanto maior o número de automóveis expostos da carteira, maior a quantidade de sinistros esperada (MARTINS, 2012).

Tabela 9: Variáveis, categorias, parâmetros e níveis descritivos dos modelos Binomial Negativa (NB) e Poisson

Variável	Categoria	Modelo Discreto	
		Parâmetro NB	Parâmetro Poisson
Intercepto		-2,2044	-1,2805
Sexo	Masculino	-	-
	Feminino e Pessoa Jurídica	-0,3715	-0,3377
Idade	Entre 18 e 25 anos	0,1143	0,3442
	Maior que 25 anos	-	-
Região	Sul	-	-
	Demais Regiões	0,6718	0,3953
Categoria Tarifária	Passeio nacional	-	-
	Demais	-0,4737	-0,9292
Ano do Modelo	< 2010	0,3638	0,1988
	≥ 2010	-	-
Expostos		0,1078	0,0318

A Tabela 10 apresenta a assertividade (percentual de classificação correta do modelo logístico) e o erro quadrático médio para os modelos ajustados. Nota-se que os modelos de Poisson Inflada de Zeros (ZIP) e Binomial Negativa Inflada de Zeros (ZINB) apresentam erro semelhante entre si e menores que os modelos das distribuições Poisson e Binomial Negativa. A assertividade, que indica o percentual de observações corretamente classificadas na primeira fase do modelo, também é semelhante entre os modelos ZIP e ZINB.

Este resultado indica que o modelo de Poisson Inflada de Zeros é muito mais assertivo para modelar o número de sinistros de automóveis da carteira de seguros brasileira no período de 2010.

Tabela 10: Erro quadrado médio de cada modelo

Modelo	Assertividade	Erro Quadrático Médio
ZINB	77,03%	1,1285
ZIP	77,44%	1,1286
Poisson	-	1.590.251
Binomial Negativa	-	12,8500

Fonte: elaborado pelo autor

5. Considerações finais

Os dados de seguro de automóveis, extraídos do sistema Autoseg da SUSEP para o segundo semestre de 2010, apresentaram grande frequência de carteiras com zero sinistros. Por esse motivo os modelos de Poisson Inflada de Zeros (ZIP) e Binomial Negativa Inflada de Zeros (ZINB) puderam ser ajustados. Ao comparar o erro quadrático médio destes modelos aos modelos clássicos (Poisson e Binomial Negativa) notou-se que as distribuições infladas de zeros apresentaram um menor erro, ou um melhor ajuste. Desta forma o primeiro objetivo deste artigo (propor um ajuste de uma distribuição mais adequada, Inflada de Zeros, à quantidade de sinistro de automóveis), foi atendido.

Os modelos ZIP e ZINB foram ajustados em duas fases. Na primeira fase estimou-se a probabilidade da não ocorrência de sinistro por meio de um modelo logístico e na segunda fase estimou-se a quantidade de sinistros esperada por meio da distribuição Poisson ou Binomial Negativa. Observou-se que, na primeira fase, os modelos logísticos apresentaram uma assertividade (percentual de observações classificadas de forma correta) em torno de 77%. Na segunda fase os modelos apresentaram um erro quadrático médio pequeno, em torno de 1,13, menor que os modelos Poisson e Binomial Negativa. Desta forma o segundo objectivo deste artigo (estimar tanto a probabilidade de ocorrência do sinistro quanto o número de sinistros por meio das distribuições propostas) foi atendido.

Os modelos ajustados consideraram dois grupos de variáveis explicativas: características dos clientes e características do ambiente, avaliando o impacto de fatores humanos (características dos condutores de veículos) e externos (características do ambiente) na quantidade de sinistros de automóveis, terceiro objetivo deste artigo. Considerando as características dos condutores, verificou-se que seguros de automóveis feitos a homens apresentam maior risco, com maior probabilidade de ocorrência de sinistro e maior média de sinistro para as carteiras destes clientes; clientes com idade de 18 a 25 anos apresentam maior risco, com maior probabilidade de ocorrência de sinistro e maior número de sinistro previsto; seguros feitos a automóveis da categoria passeio nacional tem maior risco que as demais, ou seja, maior chance de ocorrência e maior número de sinistros previsto; clientes com automóveis mais antigos, com modelos anteriores a 2010, possuem maior risco. Considerando as características do ambiente, observou-se que carteiras de sinistros de automóveis com maior quantidade de automóveis expostos possuem maior probabilidade e maior número esperado de sinistros; quanto maior a importância segurada média da carteira, ou seja, carteira

ISSN 1984-6606

caracterizada por mais carros de luxo ou mais caros, menor o risco, ou seja, menor a quantidade de sinistros esperada; automóveis da região Sul apresentam menor risco. Estes resultados são semelhantes aos encontrados por Filho e Lugon (2004), Boucher et al. (2009), Bortoluzzo et al. (2011) e MARTINS (2012).

Como limitações, este estudo considerou uma carteira de seguros do segundo semestre de 2010 e não considerou diferenças existentes dentro das regiões, exemplo, interior ou região metropolitana de São Paulo. Como sugestões futuras, pode-se considerar a divisão entre região metropolitana e demais regiões nos modelos estimados.

REFERÊNCIAS

BAILEY, R. A.; SIMON, L. J. Two studies in automobile insurance ratemaking. ASTIN **Bulletin**, v. 47, n. 1, p. 1-19, 1960.

BENNETT, M. C. Models in motor insurance. **Journal Student Society**, v. 22, n. 1, p. 134-160, 1978.

BORTOLUZZO, A.; CLARO, D.; CAETANO, M.; ARTES, R. Estimating Total Claim Size in the Auto Insurance Industry: a Comparison between Tweedie and Zero-Adjusted Inverse Gaussian Distribution. **Brazilian Administration Review**, v. 8, n. 1, p. 37-47, 2011.

BOUCHER, J.-P.; DENUIT, M.; GUILLEN, M. Number of Accidents or Number of Claims? An Approach with Zero-Inflated Poisson Models for Panel Data. **Journal of Risk and Insurance**, v. 76, n. 4, p. 821-846, dez 2009.

BÖHNING, D.; DIETZ, E.; SCHLATTMANN, P. Zero-Inflated count models and their applications in public health and social science. **Applications of Latent Trait and Latent Class Models in the Social Sciences**, v. 1, n. 32, p. 333-344, 1995.

CARVALHO, J.; LAVOR, S. C. Repeat Criminal Victimization and Income Inequality In Brazil. XXXVI Encontro Nacional de Economia. Anais... Salvador, BA: **XXXVI Encontro Nacional de Economia**, 2008.

COHEN, A. C. Estimation in Mixtures of Discrete Distributions. International Symposium on Discrete Distributions. **Montreal: International Symposium on Discrete Distributions**, 1963.

CREPON, B.; DUGUET, E. Research and development, competition and innovation pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity. **Journal of Econometrics**, v. 79, n. 2, p. 355-378, ago 1997.

DELGADO, M. Projecto solvência II - Modelação do risco de subscrição numa companhia de seguros não vida. **Universidade Nova de Lisboa**, 2011.

FERNANDEZ, P.; BOTELHO, D. Incorporação da Heterogeneidade em Modelos de Variáveis com Dados de Contagem Aplicados ao Marketing. **XXIX Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração (EnANPAD).** Brasília, DF: XXIX EnANPAD, 2005.

FILHO, H. Z.; LUGON, V. B. P. Modelo para Gestão do Risco nas Propostas de Seguro de Automóvel, com Base no Perfil Socioeconômico e Cultural do Segurado, Utilizando Análise Discriminante. **XXVIII Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração (EnANPAD)**. Curitiba, PR: XXVIII EnANPAD, 2004.

GURMU, S. Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. **Journal of Applied Econometrics**, v. 12, n. 3, p. 225-242, 1997.

GURMU, S.; TRIVEDI, P. Excess zeros in count models for recreational trips. **Journal of Business & Economic Statistics**, v. 14, n. 4, p. 469-477, 1996.

HEILBRON, D. **Generalized linear models for altered zero probabilities and overdispersion in count data.** San Francisco, 1989.

HEILBRON, D. Zero-altered and other regression models for count data with added zeros. **Biometrical Journal**, v. 36, n. 5, p. 531-547, 1994.

JOHNSON, N.; KOTZ, S. **Distributions in Statistics: Discrete Distributions**, Boston: Houghton Mifflin, Wiley/Houghton-Mifflin, 1969. 328 p.

JOHNSON, P. D.; HEY, G. B. Statistical Studies in motor insurance. **Journal of the Institute of Actuaries**, v. 97, n. 1, p. 199-249, 1971.

JUNG, J. A. N. On automobile insurance ratemaking. **ASTIN Bulletin**, v. 5, n. 1, p. 41-48, 1968.

JUNIOR, A. E. Uso da distribuição de Poisson para avaliar a evolução da taxa de ocorrência de sinistros em uma carteira. **Revista Brasileira de Estatística**, v. 64, n. 221, p. 67-79, 2003.

KHAMKONG, M. Comparing Models for Fitting Zero-inflated Data. 6th IMT-GT International Conference on Mathematics, Statistics and their Applications. Kualaq Lumpur: **6th IMT-GT International Conference on Mathematics, Statistics and their Applications**, 2010.

LAMBERT, D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. **Technometrics**, v. 34, n. 1, p. 1-14, 1992.

LAWLESS, J. F. Negative binomial and mixed Poisson regression. **The Canadian Journal of Statistics**, v. 15, n. 3, p. 209-225, 1987.

LORD, D.; WASHINGTON, S. P.; IVAN, J. N. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. **Accident Analysis and Prevention**, v. 37, n. 1, p. 35–46, 2005.

MARTIN, T. G.; WINTLE, B. A; RHODES, J. R. et al. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. **Ecology letters**, v. 8, n. 11, p. 1235-1246, nov 2005.

MARTINS, S. **Modelo de avaliação de risco em acidentes no ramo automóvel.** Universidade Nova de Lisboa, 2012.

MIAOU, S. P. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. **Accident; analysis and prevention**, v. 26, n. 4, p. 471-82, ago 1994.

SANTOS, K.; KUPEK, E.; CUNHA, J.; BLANK, V. Absenteísmo-doença, modelo demanda-controle e suporte social: um estudo caso-controle aninhado em uma coorte de trabalhadores

de hospitais, Santa Catarina, Brasil. **Revista Brasileira de Epidemologia**, v. 14, n. 4, p. 609-619, 2011.

SINGH, S. A note on zero inflated Poisson distribution. **Journal of the Indian Statistical Association**, v. 1, n. 1, p. 140-144, 1963.

SOUZA, W. F. **Estudo prospectivo do impacto da violência na saúde mental das tropas de paz brasileiras no Haiti**. Tese de doutorado em Saúde Pública. Fundação Oswaldo Cruz, Rio de Janeiro, Fundação Oswaldo Cruz, 2011.

VIAENE, S.; DEDENE, G.; DERRIG, R. Auto claim fraud detection using Bayesian learning neural networks. **Expert Systems with Applications**, v. 29, n. 3, p. 653-666, out 2005.

WANG, P. A bivariate zero-inflated negative binomial regression model for count data with excess zeros. **Economics Letters**, v. 78, n. 3, p. 373-378, mar 2003.

WELSH, A. H.; CUNNINGHAM, R. B.; DONNELLY, C. F.; LINDENMAYER, D. B. Modelling the abundance of rare species: statistical models for counts with extra zeros. **Ecological Modelling**, v. 88, n. 1, p. 297-308, 1996.

YAU, K. K. W.; WANG, K.; LEE, A. H. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. **Biometrical Journal**, v. 45, n. 4, p. 437-452, jun 2003.

YIP, K. C. H.; YAU, K. K. W. On modeling claim frequency data in general insurance with extra zeros. **Insurance: Mathematics and Economics**, v. 36, n. 2, p. 153-163, abr 2005.