# 3
# Metodologia

Conforme foi mencionado na seção 2.2.2, a classificação de risco original é formada pela combinação de três variáveis identificadoras: "*tipos de veículos semelhantes*", "*regiões de tarifação*" e "*idade dos veículos*".

Visando diminuir o impacto, ou mesmo extinguir a individualização do risco, a intenção da metodologia proposta é agrupar as classes. Porém, seja qual for o critério de agrupamento utilizado, ele não deve transgredir a duas restrições básicas: a eqüidade de taxas e a lógica de agrupamento.

A eqüidade impõe que um mesmo nível de cada uma das variáveis de classificação apresente necessariamente a mesma participação sobre a taxa, o que faz com que o agrupamento deva ser realizado separadamente sobre cada uma das componentes que, associadas, reproduzem as classes de risco.

É comum que especialistas, decisores, bem como o próprio mercado, apresentem crenças, ou opiniões, sobre a relação entre o risco e cada uma das variáveis que compõe a classificação. Por esta razão, a lógica na formação dos agrupamentos faz com que, seja qual for o método de agrupamento utilizado, ao final, e para que sua adoção se torne de interesse das companhias, os grupos devam apresentar um comportamento que corresponda às expectativas.

Considerando as restrições, para reduzir o número de classes de risco e, conseqüentemente, obter estimativas mais precisas e confiáveis na estimação das taxas, é necessário que a agregação seja realizada nas três direções. É necessário ainda, que os atributos considerados sejam capazes de representar as diferenças lógicas existentes.

A metodologia proposta visa atender as duas restrições apresentadas, bem como tornar mais eficiente, a estimação do risco.

Neste capítulo serão apresentados os argumentos e ferramentas utilizadas no desenvolvimento e validação da metodologia proposta por esta dissertação. O capítulo se divide em duas seções.

Na primeira seção será apresentada a metodologia de classificação de risco proposta. Esta metodologia, que é representada por um algoritmo capaz de reduzir

o número de classes de risco seguindo a direção de critérios pré-estabelecidos, conta com a aplicação de métodos estatísticos que serão apresentados no decorrer da seção.

Em seguida, na segunda seção, serão apresentados dois métodos de validação da metodologia de classificação proposta. Uma das formas de validar a classificação será avaliar o erro diante da lógica subjetiva, que considera o pensamento de especialistas. Outra maneira de avaliar a classificação obtida será verificar a eficiência de estimação do risco, diante das classes formadas. No caso da estimação, duas serão as medidas utilizadas para comparar as estimativas obtidas antes da classificação proposta, durante as iterações do algoritmo e na sua fase final.

## 3.1. Metodologia de classificação proposta

As principais razões que levaram ao desenvolvimento desta metodologia foram as conseqüências da classificação individualizada sobre a estimação do risco. Acredita-se que uma classificação com esta característica leve à ausência de sinistros em grande parte das classes e à geração de *outliers*[17].

Sabe-se que para obter estimativas confiáveis através de métodos estatísticos é necessário que exista volume de informações suficiente. Portanto, se esperaria que estimativas do risco para determinada classificação fossem pouco confiáveis diante da estrutura individualizada que é geralmente utilizada pelas seguradoras brasileiras.

O principal objetivo da metodologia proposta é promover transformações sobre a classificação individualizada do risco, capazes de reduzir o número de classes, produzindo assim, uma nova classificação, mais direta e eqüitativa, no que tange ao volume de informações.

Espera-se que, diante da agregação, o grau de compreensão, ora promovido pela classificação individualizada, seja inferior. Porém, espera-se também que os ganhos de precisão na estimação das taxas de risco e, conseqüentemente, das indenizações, sejam compensadores e satisfaçam aos interesses dos tomadores de decisão.

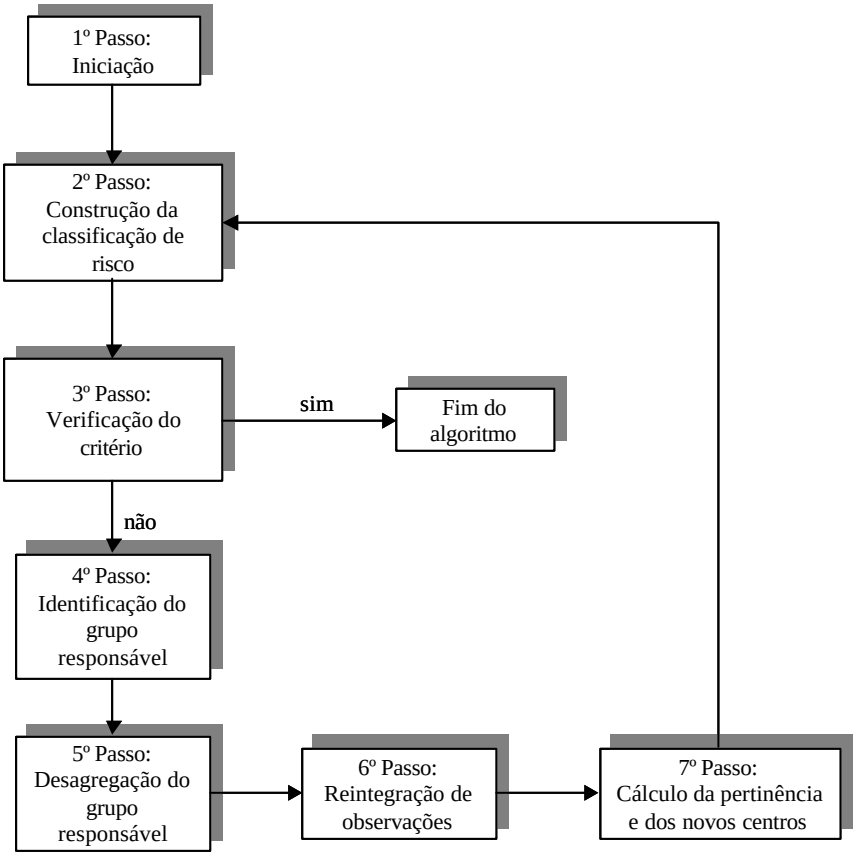
[17] Detalhes sobre o número de classes sem sinistros, bem como a geração de *outliers*, podem ser vistos na seção 2.2.1.

A metodologia proposta visa agregar as classes de risco de maneira coerente com a classificação utilizada pelas seguradoras. Portanto, está estruturada para reduzir o número de classes através da agregação realizada em cada uma das vertentes que formam a classificação original.

Considerando que grande parte das seguradoras constrói sua classificação através da combinação do *"tipo de veículo"*, com a *"região de tarifação"* e a *"idade do veículo"*, a metodologia propõe a realização do agrupamento de maneiras distintas.

A proposta, conforme mostra o diagrama abaixo, é a construção de um algoritmo de classificação, que diante de um critério determinado, seja capaz de reduzir o número de agrupamentos.

Figura 5: Diagrama de Iteração do Algoritmo de Classificação Proposto.




De acordo com o diagrama, o primeiro passo representa a iniciação do algoritmo. Nesta etapa, a proposta é de apresentar uma classificação mais agregada do que a original, porém de forma que seja possível a realização de re-

agrupamentos. Basicamente, o que se espera na iniciação é reduzir o número de iterações visando atender em menor tempo ao critério estabelecido.

No segundo passo, os agrupamentos de tipos de veículos, regiões de tarifação e idades dos veículos, que deram início ao processo, são combinados reproduzindo a primeira classificação de risco proposta.

No terceiro passo, o algoritmo verifica se a classificação proposta atende ao critério estabelecido. Em caso positivo, o algoritmo é finalizado; caso contrário, segue para o quarto passo.

No quarto passo, o algoritmo identifica, dentre os grupos de tipos, regiões e idades, qual é o maior responsável pelo não êxito da classificação proposta. É selecionado aquele grupo que apresentar maior percentual de casos em oposição ao critério.

No quinto passo, o grupo selecionado é diluído, e seus membros passam a pertencer a grupos[18] próximos.

Após o quinto passo, o algoritmo retorna ao segundo, e se inicia uma nova iteração.

Diante do algoritmo proposto, pode-se concluir que, após cada iteração, e caso o critério não tenha sido atendido, a poda de um grupo, que pode ser de veículos, regiões ou idades, é realizada. Porém, detalhes sobre o estabelecimento do critério, a iniciação (1ºpasso) e o re-agrupamento (5ºpasso), devem ser esclarecidos.

Quanto aos critérios, serão testados três tipos, sendo dois relacionados com a busca de uma classificação menos individualizada, e um visando a eqüidade entre a indenização total estimada e o risco real.

No caso da obtenção de uma classificação mais diluída, serão estabelecidos como critérios a extinção de classes sem sinistro e a extinção de classes que apresentem exposição inferior a unidade. No caso da relação entre risco real e risco estimado, o critério será a obtenção de grupos onde o erro médio percentual (MPE)[19] da classificação esteja contido no intervalo [-1%; 1%].

[18] No agrupamento, veículos se associam a grupos de veículos, regiões a grupos de regiões e idades a grupos de idade.

[19] O MPE, bem como a estimação da indenização, serão apresentados na seção 3.2.

Quanto à iniciação e ao re-agrupamento, os procedimentos adotados são semelhantes entre si, porém cada variável que compõe a classificação de risco é tratada de forma diferente.

No caso dos tipos de veículos, propõe-se a utilização de variáveis que representam o perfil dos contratos como atributos de *clusterização*[20].

As variáveis de perfil foram selecionadas por acreditar-se que, devido a representarem o perfil dos proprietários e condutores dos veículos, seja possível, por seu intermédio, identificar a relação entre a preferência por veículos e o perfil de risco dos contratos. Porém, é de interesse reduzir o número de variáveis que será utilizado no agrupamento, e entender como elas se relacionam. Com esse objetivo, inicialmente será aplicado o *Método de Análise Fatorial*.

De acordo com Jonhson e Wichern (1998), o modelo fatorial pode ser definido da seguinte forma:

Seja $X$ um vetor aleatório observável, contendo $p$ componentes, com vetor de média $m$ e matriz de covariância $\Sigma$. O modelo fatorial postula que $X$ é linearmente dependente de um pequeno número de variáveis aleatórias não observáveis $F_1,...,F_m$, chamados fatores comuns, e $p$ fontes adicionais de variação $e_1,...,e_p$, chamados erros ou fatores específicos. Desta forma, em notação matricial, o modelo de análise fatorial pode ser descrito por:

$$X_{(p\times 1)} = m_{(p\times 1)} + L_{(p\times m)}F_{(m\times 1)} + e_{(p\times 1)};$$

onde $L$ representa a matriz de carregamento dos fatores.

O que distingue o modelo de análise fatorial do modelo de regressão multivariado é o fato de que, além do vetor $e$, também a matriz $F$, que ocupa a posição das variáveis independentes, é não observada.

Sendo estas quantidades não observadas, seria inviável buscar a estimação direta do modelo fatorial a partir das observações contidas em $X$. Porém, usando certas suposições adicionais sobre $F$ e $e$, o modelo descrito implica em relações de covariância que podem ser verificadas. Além da hipótese de independência entre $F$ e $e$, as demais suposições são expressas por:

$$\mathrm{E}(F) = 0 \text{ e } Cov(F) = I;$$

$$\mathrm{E}(e) = 0 \text{ e } Cov(e) = y;$$

[20] Entender por *clusterização* o agrupamento, vice-versa.

onde $y$ é uma matriz diagonal.

Denotando por $l_{ij}$ os elementos da matriz L, o modelo fatorial ortogonal implica na seguinte estrutura de covariância para $X$:

$$Var(X_i) = \ell_{i1}^2 + ... + \ell_{im}^2 + y_i;$$

$$Cov(X_i, X_k) = \ell_{i1}\ell_{k1} + ... + \ell_{im}\ell_{km};$$

$$Cov(X_i, F_j) = \ell_{ij};$$

Um ponto de interesse da análise fatorial é conhecer a porção da variância que a i-ésima variável compartilha com as demais variáveis, o que é conhecido como comunalidade; e conhecer a porção da $Var(X_i) = s_{ii}$ dada pelos fatores específicos, o que é conhecido por variância específica.

Representando a i-ésima comunalidade por $h_i^2$ e a variância específica por $y_i$, a $Var(X_i)$ pode ser expressa como:

$$s_{ii} = \ell_{i1}^2 + ... + \ell_{im}^2 + y_i;$$

$$s_{ii} = h_i^2 + y_i;$$

Neste caso, a i-ésima comunalidade poderia ser entendida como a soma dos quadrados dos carregamentos da i-ésima variável sobre os $m$ fatores comuns.

Existem métodos aplicados na estimação dos carregamentos $\ell_{ij}$ dos $m$ fatores, e das variâncias específicas $y_i$. Os métodos das Componentes Principais e da Máxima-Verossimilhança são dois dos mais utilizados.

Para possibilitar a identificação das variáveis que apresentariam maior carga sobre determinado fator e, conseqüentemente, obter maior sensibilidade na interpretação dos fatores, são aplicadas transformações lineares sobre a matriz de carregamento $L$. Estas transformações são conhecidas como rotações de fatores. As transformações mais conhecidas e aplicadas com este intuito são a Rotação Varimax e a Rotação Oblíqua.

Por vezes é interessante substituir as variáveis pelos escores fatoriais, como são conhecidos os valores estimados para os fatores. Existem algumas técnicas que possibilitam a estimação dos escores fatoriais, e dois dos métodos mais conhecidos são o Método dos Mínimos Quadrados Ponderados e o Método de Regressão.

Na construção dos fatores que representarão as variáveis de perfil dos contratos, os carregamentos fatoriais serão estimados através do Método das

Componentes Principais. Após a estimação, pretende-se determinar o número de fatores através da variância explicada por cada fator. Como, através deste método, o primeiro fator explica mais a variância do que o segundo, e assim sucessivamente, serão considerados apenas os primeiros fatores que se apresentarem relevantes na explicação da variância total.

Como um dos objetivos da aplicação desta técnica é a compreensão da relação entre as variáveis de perfil e a interpretação do significado dos fatores, a Rotação Varimax será aplicada.

Definido o número de fatores e compreendida a relação de cada fator com cada conjunto de variáveis, os escores fatoriais serão estimados através do Método de Regressão e, finalmente, substituirão as variáveis de perfil.

Definidos os escores fatores para cada tipo de veículo, os mesmos serão utilizados no agrupamento.

O agrupamento será realizado através do algoritmo de *clusterização* denominada por Bezdek (1984) como *fuzzy c-means* (FCM)[21]. De acordo com Bezdek, o FCM pode ser definido da seguinte forma:

Seja $x_j$ um vetor contendo atributos que definem $n$ observações. O algoritmo de agrupamento *fuzzy*, conhecido como *fuzzy c-means*, é utilizado na intenção de agrupar as $n$ observações em $c$ *clusters*, considerando os atributos contidos em $x_j$.

Para a obtenção dos *clusters*, o algoritmo minimiza a seguinte função objetivo:

$$J(u_{ij}, v_k) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - v_i \right\|^2 ;$$

onde $m \in [1, \infty)$ é chamado coeficiente *fuzzy*, e $v_k$ representa o centróide do k-ésimo *cluster*.

O coeficiente *fuzzy* é responsável pelo grau de *fuzzificação* dos elementos de $x_j$.

Portanto, entendendo que $i = 1,...,c$; $j = 1,...,n$; e que $u_{ij} \in (0,1)$, o resultado *fuzzy* obtido pelo algoritmo pode ser expresso pela matriz $U = [u_{ij}]$, onde $U$

[21] Agradecimentos ao Professor Ricardo Tanscheit e a Alexandre Zanini por esclarecimentos sobre o FCM.

representa a pertinência de cada elemento $x_j$ a cada um dos $i$ *clusters* formados. Quanto maior for $m$, mais *fuzzy* se torna a matriz $U$. Caso $m$ seja igual a 1, a função objetivo $J(u_{ij}, v_k)$ é reduzida ao algoritmo de *clusterização* conhecido como *k-means.*

Minimizar a função objetivo $J(u_{ij}, v_k)$ se resume à solução das seguintes equações:

$$\frac{\partial J(u_{ij}, v_k)}{\partial u_{ij}} = 0 \quad \text{e} \quad \frac{\partial J(u_{ij}, v_k)}{\partial v_k} = 0.$$

Tal solução pode ser expressa da seguinte forma:

$$u_{ij} = \frac{\left( \frac{1}{\|x_j - v_i\|} \right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c} \left( \frac{1}{\|x_j - v_k\|} \right)^{\frac{1}{m-1}}} \quad \text{e} \quad v_i = \frac{1}{\sum_{j=1}^{n} u_{ij}^m} \sum_{j=1}^{n} u_{ij}^m x_j .$$

Com isso, o algoritmo de *clusterização fuzzy* pode ser descrito através da seguinte seqüência de passos:

1º passo: inicialização através do número $c$ de *clusters* , do critério de parada $e$, do coeficiente *fuzzy* $m$, e da matriz $U^0$ contendo os graus de pertinência iniciais.

2º passo: cálculo dos centróides dos $c$ *clusters*.

3º passo: atualização da matriz $U^{k+1}$.

4º passo: cálculo de $\nabla = \|U^{k+1} - U^k\|$. Se $\nabla < e$, fim das iterações; caso contrário, $k = k+1$ e retornar ao 2º passo do algoritmo.

Para determinar o valor do coeficiente *fuzzy*, pode-se utilizar a medida de entropia de Kullback-Liebler[22]. Esta medida pode ser expressa por:

$$KL_j = \sum_{i=1}^{c} p_i \left[ \log \left( \frac{p_i}{u_{ij}} \right) \right];$$

onde $p_i = \frac{1}{c}$ representa o caso extremo em que cada indivíduo apresenta a mesma relação de pertinência a todos os $c$ *clusters*; e $u_{ij}$ representa a verdadeira relação

[22] Cover e Joy (1991).

de pertinência de cada indivíduo a cada um dos $c$ *clusters* formados. Desta forma, $KL_j$ mede a distância entre estas duas distribuições.

Como pretende-se utilizar a medida de entropia de Kullback-Liebler para avaliar a capacidade de discriminação do algoritmo, é calculado o seu valor médio para os diversos valores do coeficiente *fuzzy* $m$. Desta forma, obtém-se que:

$$\overline{KL} = \frac{\sum_{j=1}^{n} KL_j}{n};$$

onde, quanto mais próximo de zero estiver $\overline{KL}$, menor a capacidade do algoritmo em discriminar os indivíduos em *clusters*.

Em resumo, os tipos de veículos serão agrupados através do FCM, utilizando como atributos os fatores estimados na Análise Fatorial.

O agrupamento de regiões de tarifação será realizado de maneira semelhante ao agrupamento de tipos de veículos. Porém, neste caso, não será aplicada a Análise Fatorial, já que os atributos de *clusterização* propostos são em número reduzido.

São propostos como atributos de *clusterização*, as *"freqüências de sinistros expostos"*, que são representadas pelo número médio de sinistros de perda parcial, total e roubo, por unidade exposta. Utilizando estes três atributos, o agrupamento de regiões segue a lógica do FCM.

Já no caso do agrupamento da idade dos veículos, a proposta é de que, na iniciação, eles permaneçam conforme a estrutura original, pois geralmente se encontram em número reduzido.

A metodologia proposta para agrupar idades difere das aplicadas no agrupamento de veículos e regiões. Neste caso, será utilizada a *"depreciação veicular"*. A depreciação veicular será definida da seguinte forma:

Seja a depreciação *proxi* da desvalorização dos veículos atrelada à idade. Assumindo esta relação, o cálculo da depreciação veicular será realizado em quatro etapas. Inicialmente será calculado o valor médio de cada tipo de veículo, em cada idade. O valor médio, representado por $\bar{d}_{ij}$, será calculado da seguinte forma:

$$\bar{d}_{ij} = \frac{\sum_{k=1}^{n} V_{ijk} E_{ijk}}{\sum_{k=1}^{n} E_{ijk}};$$

onde $V_{ijk}$ representa o valor de mercado do veículo $k$ que está associado ao tipo $j$ e à idade $i$; e $E_{ijk}$ representa sua exposição.

Em seguida, o valor de mercado dos tipos de veículo será padronizado em relação à sua média e ao desvio entre as idades. O valor de mercado padronizado, dado por $d_{ij}$, será calculado da seguinte forma:

$$d_{ij} = \frac{\left(\bar{d}_{ij} - \bar{\bar{d}}_{j}\right)}{Std_{j}(\bar{d}_{ij})};$$

onde $\bar{\bar{d}}_{j}$ representa o valor médio do tipo de veículo $j$, e $Std_{j}(\bar{d}_{ij})$ o desvio-padrão do valor entre as idades.

Obtido $d_{ij}$, o valor médio por idade, dado por $\bar{d}_{i}$, será calculado da seguinte forma:

$$\bar{d}_{i} = \frac{\sum_{j=1}^{l} d_{ij} E_{ij}}{\sum_{j=1}^{l} E_{ij}};$$

onde $E_{ij}$ representa a exposição de cada tipo de veículo por idade.

Finalmente, e de acordo com a seqüência de passos apresentada, a distância entre os valores médios por idade determinaria a depreciação veicular. A depreciação, representada por $\Delta_{ii'}$, será calculada da seguinte forma:

$$\Delta_{ii'} = \frac{\left|\bar{d}_{i'} - \bar{d}_{i}\right|}{\bar{d}_{i}}$$

onde $i$ representa a idade a ser agrupada, e $i'$ suas "vizinhas" mais próximas.

Acredita-se que, através do cálculo da depreciação veicular, seria possível avaliar, por exemplo, se veículos com 1 ano de fabricação encontram-se tão depreciados, que seus valores, em média, se aproximariam mais dos valores de veículos de 2 anos do que do valor de veículos novos.

Definida a metodologia de classificação, a seção seguinte apresenta as duas formas de validação de classificação propostas por esta dissertação.

## 3.2. Validação da classificação

Duas serão as formas de validar a classificação obtida através da aplicação do algoritmo proposto. Uma tem como base a lógica de agrupamento, e a outra, o erro de estimação do risco diante da classificação.

Aquela que se baseia na lógica propõe avaliar a classificação através da mensuração do erro de associação de veículos, regiões e idades, em seus respectivos agrupamentos. Neste caso, a definição de erro será subjetiva, buscando a avaliação dos grupos formados através da lógica.

No caso do agrupamento de tipos de veículos, a semelhança lógica prevalecerá. Neste caso, veículos que em nada se assemelham a outros membros de seu grupo, serão penalizados.

Em seguida, será calculado o percentual de tipos de veículos penalizados, diante dos demais tipos pertencentes ao mesmo grupo. Finalmente, a média do percentual de erro por grupo definirá o erro de classificação lógica dos tipos de veículos.

No caso do agrupamento de regiões de tarifação, a lógica geográfica será preponderante. Regiões que se apresentem distantes das demais que formam seu grupo, serão penalizadas. Neste caso, a estatística que definirá o erro de classificação lógica das regiões, será a mesma aplicada aos tipos de veículos.

Dentre todos os casos, a avaliação de agrupamento mais subjetiva será a realizada diante das idades, onde serão penalizados os pontos em que a classificação final não atende à expectativa de especialistas.

A validação da metodologia através da estimação pretende mensurar o erro, gradativamente, seguindo a evolução do algoritmo. Neste caso, será calculado inicialmente, o erro de estimação do risco diante de uma classificação originalmente individualizada. Em seguida, passo a passo, a cada iteração do algoritmo, até a sua finalização, o erro será medido.

O erro será mesurado considerando duas variáveis: a taxa de risco estimada para cada classe, e a indenização total.

A estimação da taxa risco $T$ será realizada através da sua decomposição. Sendo assim, a taxa de risco estimada será obtida da seguinte forma:

$$\hat{T}_w = \hat{S}_w \times \left( \frac{\hat{C}_w}{E_w} \right);$$

onde $w$ representa a classe de risco, $S$ representa a severidade, ou grau médio de dano, $C$ representa o número de sinistros e, finalmente, $E$ representa a exposição ao risco[23].

De acordo com a definição apresentada acima, a estimação da taxa de risco será obtida através da estimação das componentes severidade e número de sinistros[24], onde a exposição será considerada constante.

Como ambas as componentes supostamente não apresentam distribuição normal, a estimação será realizada através da aplicação da *Teoria dos Modelos Lineares Generalizados*.

Segundo McCullagh e Nelder (1989), esta teoria é definida da seguinte forma:

Seja $y$ um vetor contendo $n$ realizações de uma variável aleatória $Y$. Considere as componentes da parte aleatória de $Y$ como sendo normalmente distribuídas, independentes, com $\mathrm{E}(Y) = m$ e variância $s^2$ constante.

Considere ainda a componente sistemática representada pelas variáveis $x_1, x_2, ..., x_p$, que produzem o preditor linear $h$. Então $h$ pode ser representado por:

$$h = \sum_{j=1}^{p} x_j b_j ;$$

A ligação entre as componentes sistemática e aleatória pode ser representada da seguinte forma:

$$h = g(m_i);$$

onde $g(.)$ é chamada de função de ligação.

Sob a ótica dos modelos lineares clássicos, a componente aleatória apresenta distribuição normal, e como função de ligação, a função identidade.

A Teoria dos Modelos Lineares Generalizados permite duas extensões: primeiro para a componente aleatória, que pode possuir qualquer distribuição, desde que a mesma pertença à família de distribuições exponenciais; e, segundo,

---

[23] Embora tenha sido comprovado empiricamente que existem métodos de estimação mais eficientes (vide Chapados *et alli* (2001) e Dugas *et alli* (2003)), e esta estrutura de modelo seja um tanto questionável, este método de estimação foi selecionado por ser um dos mais conhecido no mercado segurador brasileiro.

[24] Embora esta dissertação só aborde a estimação do risco através do produto da severidade pelo número de sinistros, seria possível estimar a indenização por classe através do produto entre a indenização média e o número de sinistros.

para a função de ligação, que pode ser qualquer função monotônica e diferenciável.

É possível demonstrar que os estimadores de Máxima-Verossimilhança dos parâmetros $b$ no preditor linear $h$ podem ser obtidos através de Mínimos Quadrados Ponderados Iterativos. Neste processo, $y$ não é mais considerado como variável dependente. Em seu lugar considera-se $z$, que é a forma linearizada de $y$ através da função de ligação.

O método de estimação é o Mínimos Quadrados Ponderados Iterativos. O processo é iterativo porque tanto $z$ quanto o ponderador $W$ dependem dos valores ajustados $\hat{m}$.

Desta forma, seja $\hat{h}_0$ o cálculo corrente do preditor linear com o correspondente valor ajustado $\hat{m}_0$, derivado da função de ligação $h = g(m)$. A variável dependente é ajustada por:

$$z_0 = \hat{h}_0 + (y - \hat{m}_0)\left(\frac{dh}{dm}\right)_0;$$

onde a derivada da ligação é avaliada em $\hat{m}_0$.

O peso quadrático é definido por:

$$W_0^{-1} = \left(\frac{dh}{dm}\right)^2 V_0;$$

onde $V_0$ é a função de variância avaliada em $\hat{m}_0$. Neste ponto regride-se $z_0$ sobre os preditores $x_1, x_2, ..., x_p$ com peso $W_0$, resultando uma nova estimativa para os parâmetros $\hat{b}_1$; obtendo assim uma nova estimativa $\hat{h}_1$ do preditor linear. O processo se reinicia, até que as modificações sejam suficientemente pequenas.

Na estimação da severidade, considera-se que a componente aleatória apresenta distribuição Gama. Além disso, como estimativas negativas para a severidade são indesejáveis, será utilizada como função de ligação o logaritmo neperiano.

A equação do modelo, contendo apenas os efeitos principais, pode ser descrita como:

$$\mathrm{E}(S_{ijk}) = m_o + \log(a_i) + \log(b_j) + \log(g_k),$$

onde $a_i$, $b_j$ e $g_k$ representam os tipos de veículos, as regiões de tarifação e a idade dos veículos

Conforme proposto por Brockman e Wright (1992), e ainda por McCullagh e Nelder (1989), o número de sinistros por classe de risco será utilizado como ponderador.

Sob esta definição do modelo, não é possível considerar classes de risco que não apresentem sinistros.

No caso da estimação do número de sinistros, supõe-se que a componente aleatória seja um processo de Poisson.

Neste modelo os efeitos serão considerados como multiplicativos, ou seja, será admitido o logaritmo neperiano como função de ligação.

A parte sistemática do modelo pode ser descrita da seguinte forma:

$$\mathrm{E}\left(S_{ijk}\right) = m_o + \log(\mathrm{E}_{\mathrm{ijk}}) + \log\left(a_i\right) + \log\left(b_j\right) + \log\left(g_k\right),$$

onde os três últimos termos representam as variáveis que formam a classificação de risco, e $\mathrm{E}_{\mathrm{ijk}}$ representa a exposição por classe.

O termo subseqüente a $m_o$ é uma variável quantitativa da qual o coeficiente de regressão é conhecido, e com valor unitário. Isto decorre da suposição de que o número de sinistros é diretamente proporcional à exposição. Este termo é conhecido na literatura como variável *offset*[25].

Para avaliar a aderência dos modelos e, conseqüentemente, a eficiência da classificação de risco formada a partir da metodologia proposta, serão utilizadas duas medidas de erro: o MAPE e o MPE.

Definido o método de estimação da taxa de risco por classe $\hat{T}_w$, a indenização estimada por classe, representada por $\hat{I}_w$, será obtida da seguinte forma:

$$\hat{I}_w = \hat{T}_w \times E_w \times IS_w;$$

onde $E_w$ representa a exposição total por classe, e $IS_w$ representa a média ponderada da importância segurada, pela exposição.

[25] Para maiores esclarecimentos sobre o termo *offset*, ver McCullagh e Nelder (1989) e *SAS Technical Report* P-243 (1994).

Após a estimação de taxas e indenizações, as medidas de erro utilizadas para validar a classificação proposta, diante da estimação, serão aplicadas.

O Erro Médio Percentual Absoluto, denominado MAPE, será calculado da seguinte forma:

$$MAPE = 100 \times \frac{\sum_{w=1}^{n} \left| \frac{\left( T_w - \hat{T}_w \right)}{\hat{T}_w} \right|}{n};$$

onde $w$ representa cada uma das $n$ classes de risco, $T$ representa a taxa de risco real e $\hat{T}$ representa a taxa de risco estimada.

Sabendo que a estimação das taxas será realizada com base em classificações de risco diferentes, esta medida de erro foi selecionada por tornar possível a comparação entre diferentes modelos.

Considerando $I_w$ como a indenização real de sinistros, referentes à classe $w$, e $\hat{I}_w$ como a indenização estimada, o Erro Médio Percentual, denominado MPE, será calculado da seguinte forma:

$$MPE = 100 \times \frac{\sum_{w=1}^{n} \frac{\left( I_w - \hat{I}_w \right)}{\hat{I}_w}}{n};$$

Esta medida será utilizada com a intenção de avaliar quão próxima a indenização estimada estaria da indenização verdadeira, em média.