UNIVERSIDADE FEDERAL DE MINAS GERAIS
Especialização em Estatística – Ênfase em Mercado e Indústria

Método Estatístico de *Análise de Cluster* Aplicado aos dados de uma Associação de Proteção Veicular

Tamires Lamon Gomes Silva

Belo Horizonte
2013

Tamires Lamon Gomes Silva

Método Estatístico de *Análise de Cluster* Aplicado aos dados de uma Associação de Proteção Veicular

Monografia de Conclusão de Curso apresentada ao curso de Especialização em Estatística, da Universidade de Minas Gerais.

Orientadora: Profa. Sueli Aparecida Mingoti

Belo Horizonte
2013

RESUMO

Objetivou-se abordar o modelo de precificação do Programa de Proteção Veicular (PPV) aplicando o método estatístico de análise de C*luster* para melhor mensuração de uma variável que determina o valor a ser cobrado por uma contratação. Equiparado a precificação do PPV à precificação de seguro de automóvel, ambos têm como objetivo a mensuração do risco, onde o prêmio cobrado pelo segurador corresponde ao valor inicial do rateio cobrado pela associação e deverão garantir os resultados satisfatórios através de análise estatística do risco. A análise da precificação de um seguro é sempre algo passível de muitas discussões. Questionamentos existem sobre quais indicadores utilizar e como consolidá-los de forma a estabelecer um critério justo de avaliação de *performance*. Embora o segmento de associações do PPV esteja recente no mercado e desprovidas de conhecimentos técnicos, essas possuem informações importantes e similares às das seguradoras para realizar estudos de precificação. É nesse sentido que se apresenta este trabalho: como uma proposta a aplicação do método estatístico de análise de *cluster* aos dados de uma associação de proteção veicular e verificar como as informações dos grupos formados poderiam ser utilizadas para uma precificação diferenciada.

Palavra chaves: Precificação; *Cluster*; Programa de Proteção Veicular.

ABSTRACT

Aimed to address the pricing model Protection Program Vehicle (PPV) using the statistical method of cluster analysis to better measurement of a variable that determines the amount to be charged for a hiring. Equated with PPV pricing to pricing automobile insurance, both are aimed at measuring risk, where the premium charged by the insurer corresponds to the initial assessment levied by the Association and shall ensure the satisfactory results through statistical analysis of risk. The analysis of the pricing of insurance is always something subject of many discussions. Questions exist about which indicators to use and how to consolidate them in order to establish a fair criterion of performance evaluation. Although the segment associations PPV is latest in the market and devoid of technical knowledge, these have important information to insurers and similar studies for pricing. In this sense, this paper presents: a proposal to apply the statistical method of cluster analysis to the data of an association to protect vehicle and check the information of the groups formed could be used for a differentiated pricing.

Key word: Pricing; Cluster; Vehicle Protection Program.

Sumário

1 INTRODUÇÃO

O mercado de seguros encontra-se com tendência crescente, segundo dado do IBGE ultrapassou 3% da participação do Produto Interno Bruto Brasileiro em 2011, sendo que 11,6% do prêmio arrecadado neste mesmo ano foram do ramo automóvel, abaixo apenas do ramo saúde e pessoas (dados registrados pela SUSEP). Em paralelo, este segmento vem sofrendo grande concorrência com o surgimento das associações que oferecem o Programa de Proteção Veicular (PPV) que tem o mesmo objetivo das segurados de proteger o patrimônio contra prejuízo financeiro.

Embora o PPV e Seguradora tenham o objetivo em comum, segundo Brasil (1985 a) a principal função é restabelecer o equilíbrio financeiro abalado por um dano causado pela ocorrência de evento coberto, o PPV não possuem um órgão fiscalizador e que estabeleça regras para sua comercialização como é o caso da entidade SUSEP – Superintendência de Seguros Privados responsável pela autorização, controle e fiscalização do mercado de seguros, previdência complementar aberta, capitalização e Resseguro no Brasil. Com isso não há método previamente definido de precificação do PPV como há para o seguro.

Segundo Mano (1997), o prêmio de seguro de automóvel pode ser determinado a partir de critérios de tarifação de acordo com a estratégia da seguradora, tais como perfil do segurado, tipo do carro, região, entre outros.

Atualmente, as seguradoras utilizam como estratégia de precificação a taxação por sua experiência e experiência do mercado, considerando as variáveis de sinistros, resultado e interesse da seguradora, região de circulação, modelo e categoria do veículo entre outras variáveis. Cada seguradora tem um método de precificação do seguro de automóvel aplicando as variáveis e estudos estatísticos que determine o prêmio que satisfaz o seu resultado e evite a seleção adversa do risco.

A anti-seleção ou seleção adversa por sua vez, de acordo com a Funenseg (1996), é a crescente possibilidade de que os segurados contratarão o seguro quando o prêmio for relativamente pequeno para o risco que esta sendo coberto ou ainda pode ser definido quando “pessoas ou organizações que têm probabilidade de perda acima da média compram mais seguros do que as que têm probabilidade abaixo da média”. (RANDALL, 2000, p. 13).

Há seguradora que separa o estado em sub-regiões para assim aplicar uma taxa para cada. Essa divisão se dá somente devido a proximidades geográficas, ou seja, regiões próximas possuem uma mesma taxação.

No caso das associações de proteção veicular ativas pelo país, foi verificado através de uma pesquisa de campo que não há um método padrão que determina o preço para inclusão no programa ou se tornar associado. Existem associações que determinam uma taxa de rateio antecipado pela Unidade de Federação (UF) do associado, pela faixa de valor de veículo, pela categoria e até mesmo por modelo do veículo. Sem levar em consideração as variáveis contidas no Questionário de Avaliação de Risco (QAR), por exemplo.

Para a precificação do seguro de automóvel é importante conhecer o perfil do risco de cada segurado que terá sua cobertura. Para avaliação deste risco seguradoras utilizam o questionário de avaliação de risco (QAR), definido pela FUNENSEG (2005 c) como um conjunto de informações sobre o(s) condutor(es) habitual(is) e sobre o uso do veículo. Cada seguradora utiliza um questionário próprio. As informações obtidas através do QAR definem a taxa de risco a ser utilizada no cálculo do prêmio de seguro.

Diante das constantes comparações realizadas entre Seguradora e PPV, conforme acima descritas, ficou demonstrado que as associações surgiram recentemente e que não estão providas de técnicas atuariais e estatísticas como as seguradoras que utilizam combinação de variáveis para determinar o prêmio do seguro, porém possuem informações relevantes, como valor do custo com sinistro, tipo e perda de sinistro, região, tipo/modelo/idade do veiculo e entre outras, que permite realizar um estudo que possibilite criar taxas de rateio antecipado diferenciadas de acordo combinações de variáveis objetivando chegar a um preço que satisfaz o seu resultado e evite a seleção adversa.

Portanto, o presente trabalho tem por objetivo demonstrar a aplicação do método estatístico de *Análise de Cluster* na precificação da proteção veicular. O propósito é agrupar uma dada região com índices semelhantes, aplicando o método estatístico de *Cluster*.

Para que o objetivo principal seja executado é necessário aplicar o método de *Cluster* para segmentar as cidades do estado do Rio de Janeiro, estudar os grupos criados, definindo o motivo de cada segmentação.

A associação do PPV, nomeada XX, diferencia a taxa de rateio para os veículos da categoria passeio de acordo com a UF do associado. Essa divisão se dá ao volume de sinistro e resultado de cada estado.

Pretende-se, portanto, aplicar o método de *Análise de Cluster* a fim de obter novas sub-regiões, conforme o mercado segurador, e analisar o agrupamento obtido de acordo com as variáveis determinadas.

Para que seja demonstrado a importância do presente estudo, descreveremos as características das associações em geral e especificas de proteção veicular e método estatístico a ser aplicado (*cluster*).

A partir dos argumentos descritos acima, questiona-se: ao aplicar o método *de Análise de Cluster* sobre os dados da associação de proteção veicular como ficaria a regionalização do estado do Rio de Janeiro? Há diferença em relação o agrupamento das cidades do Rio de Janeiro por proximidade geográfica, conforme utilizado por algumas seguradoras, e a solução obtida pelo método de *Cluster*?

2 ASSOCIATIVISMO E COOPERATIVOS

Segundo ABRANTES (2004), o associativismo é um sistema privado, sem fins lucrativos, que tem por objetivo a defesa e promoção dos interesses das pessoas (físicas e/ou jurídicas). Tal sistema pressupõe a constituição de organizações, denominadas Associações, sendo que tais organizações, podem implementar programas de benefícios em geral visando contemplar os interesses e necessidades de seus associados e da comunidade em geral, em estrita observância da legislação pertinente e de seu estatuto social.

O associativismo surgiu, de acordo com Gasparini (2010), já nos primórdios da humanidade, quando o homem percebeu a necessidade de viver em grupos para caçar, se defender e cultivar. Na era industrial foi obrigado a se organizar mais para enfrentar as condições precárias de trabalho e na era atual, a era do conhecimento, é necessário buscar o desenvolvimento econômico e social através de grupos estruturados e preparados.

Segundo pesquisas de diversos autores as associações beneficentes surgiram no Brasil no século XIX, já na década de 1830 há registros da presença de sociedades de socorros mútuos. Mas seu verdadeiro crescimento se deu na década de 1890, especialmente no Rio de Janeiro e em São Paulo. O estado de Minas Gerais viu proliferar o movimento mutualista apenas na década de 1910, embora houvesse tais organizações desde a década de 1870.

A título de noticiário, segundo o Jornal do Comércio (10/2010), as cooperativas existentes no Rio Grande do Sul, em vários setores de atividades, como agricultura, alimentação, finanças, saúde, comercialização, seguros e crédito, têm um faturamento superior a R$ 27 bilhões, o que representa 11,3% do Produto Interno Bruto do Estado, e estão em processo de crescimento. Só de 2010 para 2011, cresceram 25,2%, contribuindo com mais R$ 5,4 bilhões. O modelo de negócio cooperativo é um fator importante no desenvolvimento econômico e social, está presente em mais de 100 países e soma mais de 800 milhões de cooperados em todo o mundo, gerando mais de 100 milhões de empregos. No Brasil, há mais de 6.650 cooperativas, com mais de 9 milhões de cooperados.

2.1 Associações de Proteção Veicular

De acordo com o SINCOR-MG, as primeiras operações de associações de proteção veicular tiveram origem em associações que agregavam caminhões e que, segundo seus fundadores, pretendiam proteger aqueles riscos renegados pelas companhias seguradoras ou aceitáveis mediante taxação altamente agravada. Com a descoberta deste nicho, ocorreram por volta do ano de 2005 os primeiros registros de associações criadas unicamente para operar o Programa de Proteção Veicular, congregando também veículos leves e motos, com sede preponderante no estado de Minas Gerais, onde se desenvolveram e solidificaram suas operações passando a disputar abertamente o mercado de consumo de seguros.

O objetivo das associações ou cooperativas de proteção veicular é similar aos serviços oferecidos pelas Seguradoras, segundo o SINCOR-MG, benefício que garante aos associados à reparação de danos ocorridos em seus veículos, quando decorrentes de colisão, incêndio, roubo e furto.

Similar ao conceito de seguros quanto às coberturas, determina a Funenseg (2001) as garantias principais no seguro de automóvel são as coberturas básicas e adicionais:

- As coberturas básicas são as ligadas diretamente ao veículo como cobertura compreensiva (colisão, incêndio e roubo), incêndio e roubo, colisão e incêndio, responsabilidade civil facultativa de veículos, acidentes pessoais de passageiros. (FUNENSEG, 2001)
- As coberturas adicionais são utilizadas como complemento a cobertura básica. As coberturas adicionais são: acessórios, carrocerias, equipamentos, assistência 24 horas, carro reserva, despesas extraordinárias,diária por perda de faturamento, extensão de perímetro, valor de novo e vidro protegido. (FUNENSEG, 2001)

2.2 Precificação do Programa Proteção Veicular

No conceito de seguro, Souza (2007) afirma que o seguro baseia no compartilhamento de risco em que a seguradora cobra um prêmio para compensar o segurado de um prejuízo, no caso de automóvel a questão que se aplica é quanto se

cobrar do segurado para dá-lo esta segurança, para que seja suficiente a seguradora para cobrir eventuais sinistros.

Não obstante, os associados ao Programa de Proteção Veicular compartilham o mesmo objetivo de se prevenir de um evento futuro e incerto proveniente de danos causado ao veículo. Porém, no conceito de associação é definido o rateio como sendo a divisão proporcional dos prejuízos apurados no mês pela quantidade de veículos ativos naquele mesmo mês. Desta forma, várias são as formulas de rateia o prejuízo adotadas pelas associações.

Diante dos resultados encontrados através de uma pesquisa de campo realizada no período de agosto/2012 a janeiro/2013, verificam-se a seguir algumas formas de ratear o prejuízo:

- Rateio Simples: somam-se todas as despesas referentes aos sinistros e divide-se pelo total de veículos ativos na associação;

- Rateio Antecipado: define-se uma taxa básica a ser aplicada ao valor FIPE (trata-se de uma tabela de referência de valor de mercado de veículos), do veículo no momento da adesão ao Programa.

Tanto o rateio simples quanto o rateio antecipado, pode ser diferenciado pela categoria tarifária, pelo Estado do Associado ou pela marca e modelo do veículo.

A categoria tarifária é determinada de acordo com o tipo do veículo, se nacional, importado, passeio, pick-up, utilitário entre outras categorias.

No âmbito do seguro, a importância dada à mensuração de um risco vem desde a época das grandes navegações, onde o preço do seguro da carga dependia do navio, não havia um calculo exato, hoje existe uma grande preocupação quanto a uma boa precificação principalmente devido à competitividade entre as seguradoras é o que informa Souza (2007).

Esta mesma preocupação na mensuração do risco está nascendo para as associações. Visto que não se pode tratar todo o risco igualmente, pois geraria a anti-seleção do risco, onde seriam atraídos “maus” riscos, afirma FUNENSEG (2001).

Em comparação ao seguro tradicional, existe a franquia que no Programa de proteção veicular é conhecido como cota de participação, porém com o mesmo conceito de Seguradora. De acordo com as informações do Tudo Sobre Seguros (2013), franquia é

uma parte da indenização que o segurado assume como responsabilidade de arcar caso haja um sinistro.

2.3 Análise Multivariada

Em quase todas as áreas de pesquisa várias variáveis são mensuradas e, em geral, essas devem ser analisadas conjuntamente. A análise multivariada é a área da estatística que trata desse tipo de estudo e existem várias técnicas que podem ser aplicadas, sendo que, a utilização dessas depende do tipo de dado que se deseja analisar e dos objetivos do estudo.

Segundo Anderson (1984), existe basicamente, duas formas de classificar as técnicas de análise multivariada: as que permitem extrair informações a respeito da independência entre as variáveis que caracterizam cada elemento, tais como análise fatorial, análise de agrupamento, análise canônica, análise de ordenamento multidimensional e análise de componentes principais; e as que permitem extrair informações a respeito da dependência entre uma ou mais variáveis ou uma com relação à outra, tais como análise de regressão multivariada, análise de contingência múltipla, análise discriminante e análise de variância multivariada.

2.3.1 Análise de Agrupamento – Cluster

A análise de agrupamentos também conhecida por outros nomes, como, *análise de cluster* ou análise de conglomerados, e dependendo da área de estudo em que é aplicada possui ainda outras denominações, é um método estatístico que permite agrupar elementos, indivíduos, produtos e até mesmo comportamentos de elementos de uma amostra, com base nas similaridades e diferenças das características que estes itens possuem (CORRAR; PAULO E DIAS FILHO 2007).

Mingoti (2005) acrescenta que o método *Cluster* é um método exploratório no qual se objetiva dividir em grupos uma população (ou amostra) sendo que em muitos casos o

número de grupos não é conhecido *à priori*, mas precisa ser estimado via os dados amostrais observados. Busca agrupar elementos amostrais baseando-se na similaridade entre eles. Os grupos são determinados de forma a obter-se homogeneidade dentro dos grupos e heterogeneidade entre eles.

Segundo Corrar, Paulo e Dias Filho (2007), se a análise de agrupamentos for usada apropriadamente ela pode acrescentar muitas informações que poderiam não ser descobertas por outros meios, atendendo dessa forma a necessidade fundamental de determinadas pesquisas.

Na análise de agrupamento, é fundamental ter particular cuidado na seleção das variáveis de partida (mensuradas) que vão caracterizar cada elemento (objeto), e determinar, em última instância, qual o grupo em que esse deve ser inscrito. Nesta análise não existe qualquer tipo de dependência entre as variáveis, isto é, os grupos se configuram por si mesmo sem necessidade de ser definida uma relação causal entre as variáveis utilizadas, sobretudo gerar hipóteses, mais do que testá-las, sendo necessário a validação posterior dos resultados encontrados através da aplicação de outros métodos estatísticos (REIS, 1997).

Genericamente, a análise de agrupamento compreende cinco etapas (Aaker et al., 2001):

1. A seleção de elementos ou de uma amostra de elementos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais serão obtidas informações necessárias ao agrupamento dos elementos;
3. A definição de uma medida de semelhança ou distância entre os elementos;
4. A escolha de um algoritmo estatístico de partição/classificação;
5. Por último, a validação dos resultados encontrados.

2.3.3 Medidas de Similaridade

Segundo Aaker et al. (2001), a premissa mais importante da análise de agrupamento é a de que a medida de similaridade ou dissimilaridade na qual o processo

de agrupamento se baseia é uma medida válida de similaridade ou dissimilaridade entre os elementos.

Pode-se definir similaridade como “a medida de correspondência, ou semelhança, entre os objetos a serem agrupados.” (CORRAR; PAULO E DIAS FILHO, 2007, p.333)

Segundo Mingoti (2005) é necessário pré-especificar a medida de similaridade a ser utilizada no agrupamento, pois existem várias medidas de similaridades diferentes sendo que cada uma delas produz um determinado tipo de agrupamento.

A maioria dos métodos de análise de *cluster* requer uma medida de similaridade entre os elementos a serem agrupados, normalmente expressos como uma função distância ou métrica.

Segundo Cormack (1971) as distâncias mais utilizadas em análise de agrupamento são:

1) Distância Euclidiana: a distância entre dois casos (i e j), é a raiz quadrada do somatório dos quadrados das diferenças entre valores de i e j para todas as variáveis (v = 1, 2, ,,,,, p).

$$d_{ij} = \sqrt{\sum_{v=1}^{p} (X_{iv} - X_{jv})^2}$$

sendo:

X_{iv} representa o valor da variável v do elemento i,

X_{jv} representa o valor da variável v do elemento j,

p é o número variáveis

2) Distância Euclidiana ao Quadrado: a distância entre dois casos (i e j), é definida como o somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis (v = 1, 2.....p).

$$d_{ij}^2 = \sum_{v=1}^{p} (X_{iv} - X_{jv})^2$$

sendo:X_{iv} representa o valor da variável v do elemento i,

X_{jv} representa o valor da variável v do elemento j,

p é o número variáveis

Define-se o nível de similaridade como:

$$S_{il} = \left(1 - \frac{d_{il}}{max\{d_{ik}, j, k = 1,2, \ldots, n\}}\right) x\ 100$$

onde, max{d_{jk}, j, k = 1,2,...,n} é a maior distância entre os e n elementos amostrais na matriz de distância Dnxm do primeiro estágio do processo de agrupamento (MINGOTI, 2005).

2.3.4 Método de Agrupamento Hierárquico

O método hierárquico de *Cluster*, segundo Bussab (1990), consiste em uma série de sucessivos agrupamentos ou sucessivas divisões de elementos, onde os elementos são agregados ou desagregados. Os métodos hierárquicos são subdivididos em métodos aglomerativos e divisivos. Os grupos, nos métodos hierárquicos, são geralmente representados por um diagrama bi-dimensional chamado de dendograma ou diagrama de árvore. Neste diagrama, cada ramo representa um elemento, enquanto a raiz representa o agrupamento de todos os elementos. A Figura 1 demostra um exemplo de dendograma.

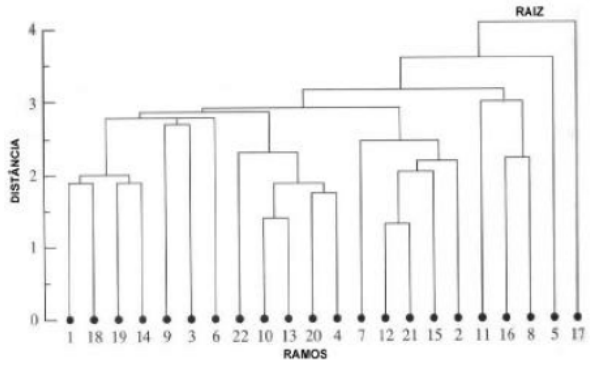


Figura 1: Denograma

De acordo com Mingoti (2005), no método aglomerativo, cada elemento inicia-se representando um grupo, e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade, até o último passo, onde é formado um grupo único com todos os elementos.

Existe uma variedade de métodos aglomerativos, que são caracterizados de acordo com o critério utilizado para definir as distâncias entre grupos. Entretanto, a maioria dos métodos são basicamente formulações alternativas de três grandes conceitos de agrupamento aglomerativo (ANDERBERG, 1973):

1) Métodos de ligação single linkage (ligação simples), complete linkage (ligação completa), average linkage (ligação das médias), median linkage (ligação de medianas);
2) Método de centróide;
3) Métodos de minimização da soma de erros quadráticos ou variâncias (método de Ward).

Neste estudo será aplicado o método de agrupamento de Ward utilizando-se a distância euclidiana ao quadrado como medida de dissimilaridade. Portanto, no texto deste trabalho aprofunda-se apenas no método Ward de agrupamento.

Ward (1963) propõe um processo geral de classificação em que n elementos são progressivamente reunidos dentro de grupos através da minimização de uma função objetivo para cada (n -2) passos de fusão. Inicialmente, este algoritmo admite que cada um dos elementos se constituía em um único agrupamento. Em cada passo do agrupamento dois grupos são unidos com base no valor da soma de quadrados dentro dos grupos. Basicamente essa medida quantifica a variabilidade dos elementos alocados em um mesmo grupo em relação ao vetor de médias do grupo. Cada grupo da partição tem a sua soma de quadrados e a adição das somas de quadrados de todos os grupos dá origem a soma de quadrados total dentro dos grupos da partição. Para cada possibilidade de agrupamento é calculado a soma de quadrados total dentro dos grupos, sendo unidos os grupos que resultam no menor valor numérico dessa soma de quadrados. Dessa forma, o método de agrupamento busca a partição que minimiza a variabilidade dentro dos grupos formados, daí ser conhecido como método de mínima variância. Pode ser

mostrado (Ward, 1963), que esse critério de agrupamento é equivalente ao agrupamento formado utilizando-se a distância definida em (1) para comparação dos conglomerados em cada passo do agrupamento, sendo unidos sempre os dois grupos que geram o menor valor numérico de (1).

$$d(C_l, C_i) = \left[\frac{n_l\, n_i}{n_l + n_i}\right] (\overline{X}_{l\,.} - \overline{X}_{i\,.})' \; (\overline{X}_{l\,.} - \overline{X}_{i\,.}) \qquad (1)$$

sendo:

n_i o número de elementos no conglomerado C_i,

n_l o número de elementos no conglomerado C_l,

$\overline{X}_{i\,.}$ o centróide do conglomerado C_i

$\overline{X}_{l\,.}$ o centróide do conglomerado C_l

C_i e C_l os conglomerados que estão sendo comparados.

2.3.5 Método de Agrupamento Não-Hierárquico

Ao contrário do método hierárquico, o método não-hierárquico de agrupamento não produz “árvore (ou dendograma)”, para demonstrar o resultado do agrupamento feito em cada passo. Não há propriedade de hierarquia, ou seja grupos unidos num determinado passo podem se separar em passos posteriores. No entanto, para execução do algoritmo é necessário que a quantidade de grupos (k) deve ser pré-estabelecida.

A partição dos dados se dá respeitando duas premissas: a coesão interna e o isolamento dos grupos é o que informa Mingoti (2005). Existem vários métodos não-hierárquicos mas para o estudo deste presente trabalho o método K-médias (Everitt et. al, 2001), foi utilizado apenas como validação do da partição escolhida pelo método hierárquico.

O método das K-médias é um dos mais conhecidos e utilizados em casos práticos, segundo Hartingan (1979) e citado por Mingoti (2007). É um método iterativo no qual em cada passo do algoritmo cada elemento da amostra é comparado com o vetor de médias do grupo (sementes do grupo), sendo alocado no grupo cuja distância é menor. Em cada passo os vetores de médias são re-calculados usando os elementos amostrais

que foram alocados aos grupos no passo anterior. A execução do algoritmo é interrompida quando não há na re-alocação dos elementos nos vários grupos formados previamente. Para a inicialização do algoritmo é necessário definir as sementes (vetores) que definem o perfil inicial de cada grupo.

2.3.6 Definição da quantidade de grupos

Determinar o número de grupos para uma base de dados é uma das tarefas mais delicadas no processamento de agrupamento.

Para Barroso & Artes (2003), o número de grupos pode ser definido a priori, através de algum conhecimento que se tenha sobre os dados, pela conveniência do pesquisador, por simplicidade, ou ainda pode ser definido a posteriori com base nos resultados da análise.

De acordo com Aaker et al., (2001), para determinar o número apropriado de grupos, existem diversas abordagens possíveis: (i), o pesquisador pode especificar antecipadamente o número de grupos (*clusters)*. Talvez, por motivos teóricos e lógicos, esse número seja conhecido. O pesquisador pode também, ter razões práticas para estabelecer o número de grupos, com base no uso que pretende fazer da partição final; (ii) o pesquisador pode estimar o número de grupos a partir do uso de um método de agrupamento hierárquico. Nesse caso, será necessário especificar algum critério para determinar o momento (passo) de interrupção ao algoritmo e conseqüente determinação do número de *clusters*. As distâncias entre os *clusters* que vão sendo formados em cada passo do algoritmo de agrupamento podem servir de guia, e o pesquisador pode escolher interromper o processo quando as distâncias excederem um valor pré- estabelecido; (iii) outra abordagem é representar, graficamente, a razão entre a variância total interna dos grupos e a variância entre os grupos, em relação ao número de grupos formados. O ponto em que surgir uma curva acentuada, um ponto de inflexão, seria a indicação do número adequado de *clusters*. Aumentar esse número além desse ponto seria inútil, e diminuí-lo seria correr o risco de misturar objetos diferentes. Existem outras medidas que podem ser

usadas para comparação de partições como o coeficiente de correlação intra-classe e a estatística Pseudo-F, dentre outros.

O coeficiente correlação de intra-classe R^2, representa a proporção da variabilidade total explicada pela partição em g* grupos feita nos dados. Quanto maior for o valor desse coeficiente, maior será a soma de quadrados entre os grupos e menor será o valor da soma de quadrados residual (dentro dos grupos). (MINGOTI, 2005)

Temos que:

$$R^2 = \frac{SSB}{SSTc}$$

sendo:

SST_c: Soma de Quadrados Total corrigida para média global em cada variável;

SSB: Soma de Quadrados Total entre os g* grupos da partição, construída no passo respectivo do algoritmo. O valor do R^2 pode ser calculado em cada passo do algoritmo de agrupamento. Seus valores devem ser usados como um critério adicional para determinação do número de grupos o coeficiente de correlação intra-classe cresce naturalmente com o aumento do número de grupos assumindo o valor máximo para o caso em que g*=n, ou seja, cada elemento do conjunto de dados é um grupo isolado.

Outro critério que pode ser usado para estimar o número g de *Clusters* da partição final é a estatística Pseudo F. Segundo Calinski e Harabasz (1974) e citado por Mingoti (2007), se F é monotonimente crescente com g*, os dados sugerem que não existe qualquer estrutura natural de partição de dados. Se, no entanto, isso não ocorrer e a função *Pseudo F* apresentar um valor máximo, o número de conglomerados e a partição referente a esse valor máximo corresponderá a “partição ideal” dos dados.

A estatística *Pseudo F* é calculada pela fórmula a seguir em cada passo do algoritmo de agrupamento:

$$PF = \frac{SSB\ /\ (g^* - 1)}{SSR/(n - g^*)} = \left(\frac{n - g^*}{g^* - 1}\right)\left(\frac{R^2}{1 - R^2}\right)$$

onde:

g é o número de grupos relacionado com a partição do respectivo passo de agrupamento;

n tamanho da amostra;R^2 coeficiente de correlação infra-classe

Alguns *softwares* estatísticos fazem automaticamente o cálculo do coeficiente de correlação intra-classe em cada passo do algoritmo de agrupamento. Entretanto, esse não é o caso do *software Minitab* for Windows. Mingoti (2005), apresenta uma estratégia para determinar o número de Grupos usando o *software Minitab*:

1. Faça o agrupamento escolhendo o número de grupos (Clusters) igual a 1. O *Minitab* irá mostrar todo o histórico de agrupamento desde o primeiro passo do algoritmo até o último;
2. O valor da Soma de Quadrados é apresentado pelo *Minitab*. Esta representa a soma de quadrados do último passo do algoritmo de agrupamento, ou seja é a Soma de Quadrados Total (SSTc);
3. Observe que o decaimento do nível de similaridade de um passo do algoritmo para outro. Escolha um nível de similaridade satisfatório. Veja o número de grupos a ele associado (k).
4. Entre no *Minitab* e peça para gerar o agrupamento considerando o valor k de números de grupos;
5. Observe o valor da Soma de Quadrados que é apresentada pelo *Minitab* na saída dessa nova análise. Esta representa a soma de quadrados relativo a partição dos dados no número de grupos k escolhido em (4), ou seja é a Soma de Quadrados dentro dos grupos formados (SSW).

Desse modo, o usuário terá condições de calcular tanto o coeficiente de correlação intra-classe quanto o valor da estatística Pseudo-F.

Qualquer que seja a abordagem empregada, é aconselhável observar o padrão total dos grupos construídos (partição). Isto pode proporcionar uma medida da qualidade

do processo de agrupamento e do número de grupos que emergem nos vários níveis do método de agrupamento utilizado.

Outro procedimento utilizado como um complemento para avaliação da estimativa do número de grupos é o da comparação dos resultados obtidos por vários métodos diferentes de agrupamento. Tendo-se um valor estimado para o número de grupos, os dados são submetidos a vários métodos hierárquicos de agrupamento e poder-se-á aferir o grau de convergência entre os vários métodos de agrupamento através de uma tabela de contingência, indicando o número de observações que se agrupam no mesmo *cluster*, entre os vários métodos, considerando-se o mesmo número de grupos. Desta forma é possível verificar a maior ou menor estabilidade das soluções encontradas, de maneira a concluir acerca da qualidade do agrupamento efetuado.

3 BASE DE DADOS

A partir da base de dados fornecida pela associação de proteção veicular, foi identificada e analisada uma região que possui maior massa de itens expostos na categoria de veículos leves passeio.

Desta forma, o Estado do Rio de Janeiro é a região a ser estudada por apresentar maior quantidade de itens expostos, considerando como informações principais da base o número de itens, prêmio e sinistros por cidade.

A base foi gerada considerando os veículos leves da categoria passeio, os sinistros de causa ocorrida como colisão, furto/ roubo e incêndio, no período de janeiro de 2011 a dezembro de 2012.

A base contém os seguintes variáveis:

- Código de identificação do associado (número da matricula);
- Data de matrícula;
- Tipo de categoria do veículo;
- Ano modelo do veículo;
- Marca do veículo;
- Modelo do veículo;
- Descrição da cidade;
- Número do sinistro;
- Causa do sinistro;
- Data de ocorrência do sinistro;
- Valor do sinistro indenizável;
- Valor do prêmio pago;
- Quantidade de sinistro por cidade do Rio de Janeiro;
- Quantidade de itens exposto por cidade do Rio de Janeiro.

3.1 Tratamento dos dados

A primeira etapa do tratamento da base de dados foi a verificação da consistência dos dados. Para isso as seguintes etapas foram executadas:

- Para cada sinistro ocorrido, verificou-se se havia o valor de indenização, e se o dado não estava zerado;
- Verificou-se se existiam dados duplicados no campo de matrícula do associado;
- Verificou-se a se existiam sinistros fora do período analisado.

Diante das informações contidas na base foram geradas as variáveis usadas para aplicação do método de *análise de Cluster* descritas a seguir*:*

1. Idade média dos veículos por cidade;

2. Índice de Frequência de sinistro por cidade, calculado pela fórmula:

$$F: \frac{\text{Quantidade de sinistro}}{\text{Quantidade de itens vigentes}}$$

3. Índice de Sinistralidade, calculado pela fórmula:

$$IS: \frac{\text{Valor de sinistro} - \alpha(\text{Valor do Sinistro})}{\text{Valor do prêmio}}$$

Sendo α o percentual de estimativa de salvados, sendo que salvados são objetos que se consegue resgatar de um sinistro e que ainda possuem valor econômico. Assim são considerados tanto os bens que tenham ficado em perfeito estado como os parcialmente danificados pelos efeitos do sinistro. No caso de um sinistro de veículo, o próprio veículo ou parte do mesmo encontrado após o pagamento de indenização por roubo ou furto total. Refere-se também ao que restou de um

veículo após o acidente indenizável pela seguradora. No entanto pode ser entender como salvados a reversão do que foi recuperado de um veículo em função de um sinistro para receita para seguradora. (Fonte: Caderno Tudo Sobre Seguro). Neste estudo foram considerados 10% de salvados.

4. Índice de Produção por cidade (IP). Este índice mede o volume de itens exposto em cada cidade. Auxilia na análise do resultado das variáveis 1, 2 e 3, citadas anteriormente, visto que se a produção for baixa os índices de frequência e sinistralidade podem não ter relevância. Entende-se como índice de produção a proporção de itens em determinada cidade em relação à quantidade de itens de toda frota exposta.

5. Índice de veículos populares (IVP). Este índice indica a proporção de veículos populares em cada cidade e aponta uma semelhança da frota em determinada região.

A partir de uma pesquisa realizada na revista Quadro Rodas realizada sobre os veículos mais vendidos e o boletim estatístico da SUSEP foi determinado que os modelos Palio, Gol, Siena, Corsa, Fiesta, Uno e Ká são considerados veículos populares e foram então os modelos utilizados neste estudo.

Para tratamento da base de dados e realização dos estudos foi utilizado o *Software* estatístico *MINITAB for Windows* versão 15, que de acordo com o site oficial do aplicativo tem por finalidade transformar dados em informações através de aplicações analíticas.

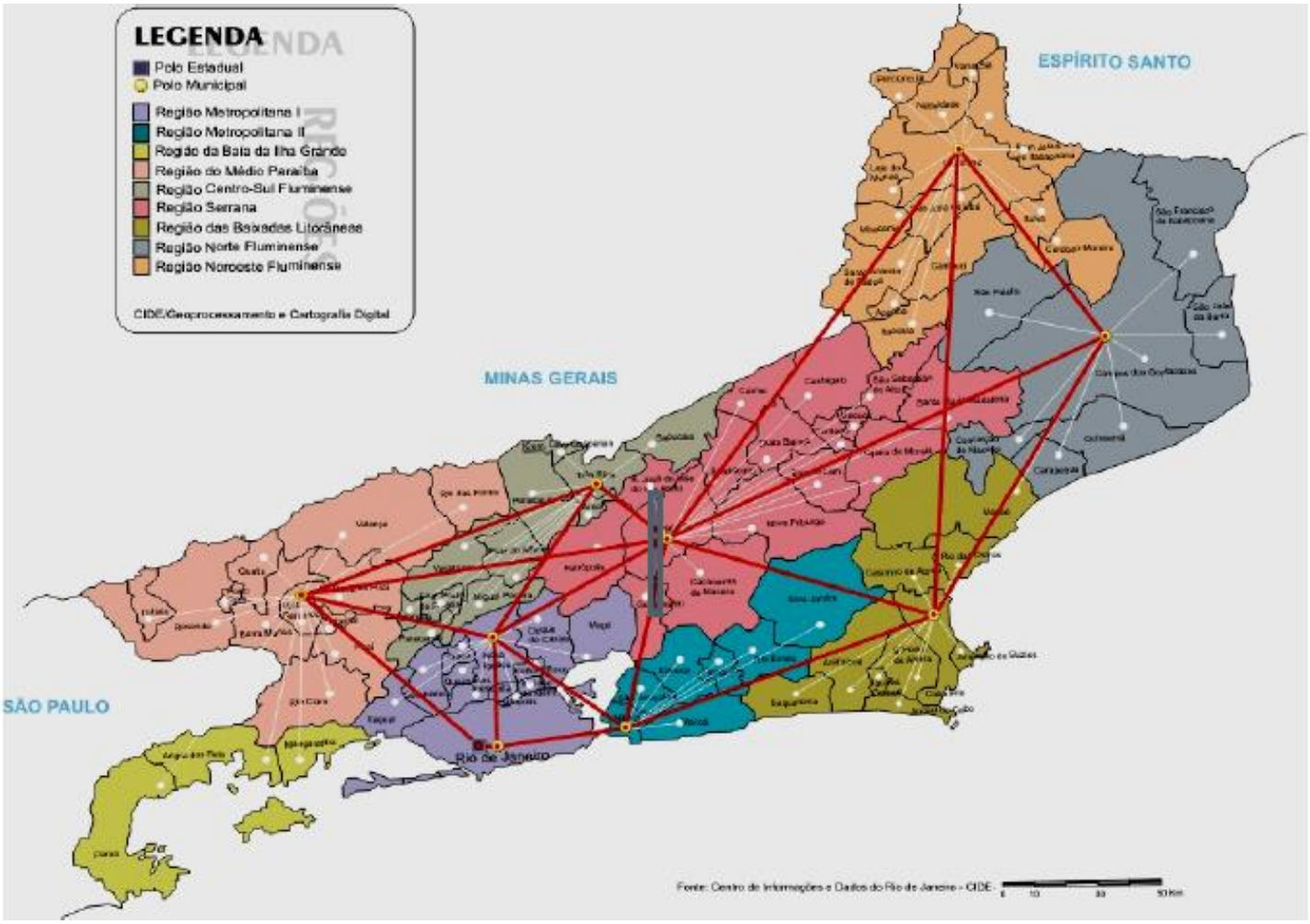
3.2 Agrupamento Utilizado Atualmente pela Associação de Proteção Veicular

De acordo com as referências de mercado algumas seguradoras utilizam o método de agrupamento por região. Este método considera a proximidade geográfica entre os bairros das cidades e provavelmente será ajustada de acordo com o comportamento da

região. Já em algumas associações o agrupamento é simplificado realizado por Unidade Federativa de acordo com o volume de itens/produção da frota de veículos.

Nesta monografia será utilizada como base de comparação a regionalização geográfica definida pelo Governo do Estado do Rio de Janeiro. Será utilizada essa informação para analisar quais as regiões das cidades formadas no agrupamento, ou seja, será analisado se os grupos formados respeitam a prática de proximidade geográfica.

As regiões do Rio de Janeiro são separadas e denominadas em: Região Metropolitana, Baía da Ilha Grande, Médio Paraíba, Centro-Sul Fluminense, Serrana, Baixadas Litorâneas, Norte Fluminense e Noroeste Fluminense (ver Figura 2).



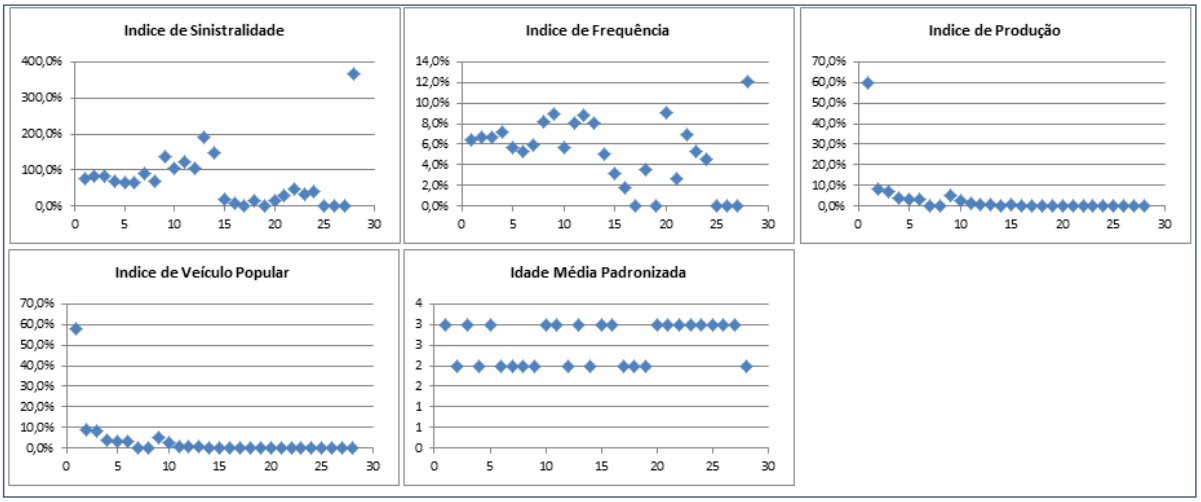
Fonte: Governo do Estado do Rio de Janeiro

Figura 2: Regionalização do Estado do Rio de Janeiro

3.3 Delineamento da pesquisa

Diante dos dados tratados verificou se que das 68 cidades do Rio de Janeiro que a associação comercializa proteção automotiva aproximadamente 41% das cidades tem volume de dados significativos para serem estudados de forma a não distorcer a análise. Portanto, as cidades que apresentam índice de produção inferior a 0,05% foram agrupadas em um pré-grupo e a análise de *Cluster* foi aplicada nas demais cidades.

Como parte do delineamento da pesquisa, foi gerado o gráfico de dispersão, apresentado na Figura 3, para análise de pontos discrepantes ou *outliers*. Foram observados que os pontos que se destacam são o Índice de Produção e o Índice de Veículo Popular, estes pontos representam a cidade do Rio de Janeiro onde se concentra o maior volume de itens expostos do Estado do Rio de Janeiro. Por este motivo Rio de Janeiro é considerada a cidade com maior importância e não pode ser retirada do estudo.



Fonte: Dados de Pesquisa

Figura 3: Gráfico de dispersão das variáveis

Um último ponto a ser avaliado antes da aplicação da Análise de *Cluster* é a a avaliação dos resultados da estatística descritiva dos dados, apresentados na Tabela 1.

TABELA 1: Estatística Descritiva das variáveis.

Variável	N	Média	Desvio-Padrão.	Mínimo	Mediana	Máximo
IS	28	0,7100	0,7710	0,0000	0,6550	3,6570
F	28	0,0521	0,0328	0,0000	0,0569	0,1207
Idade_Média	**28**	**6,8550**	**0,8060**	**4,9090**	**6,7410**	**8,7270**
IP	28	0,0355	0,1122	0,0005	0,0033	0,5961
IVP	28	0,0355	0,1095	0,0001	0,0034	0,5801

Com base na Tabela 1, o índice de sinistralidade (IS), que mede quanto do prêmio recebido está sendo direcionado para despesa de sinistro, é em média 71% entre as cidades do estado do Rio de Janeiro que possuem produção acima de 0,05%.

Índice de veículos populares (IVP) está entre 0,01% até 58,01% indica que entre as cidades analisadas todas possuem modelos de veículos classificados como popular.

Ao observar a variável Idade Média, verifica-se que a Associação de Proteção Veicular em estudo tem uma base de veículos relativamente novos no Rio de Janeiro, esses possuem idade média de 07 anos aproximadamente, sendo a idade média mínima igual 05 anos.

A dispersão dos dados em relação ao valor esperado (média) da variável é baixo em todas as variáveis observadas, exceto na idade média e Índice de Sinistralidade (IS), porém a variável Idade Média é a única que está em escala maior de unidade em relação as demais variáveis, sendo assim decidiu-se padronizá-la.

Segundo Corrar, Paulo e Dias Filho (2007) a padronização através das variáveis é uma forma comum em que se converte cada variável em escores padrões, que são obtidos pela subtração do valor de cada variável pela respectiva média e dividindo-se o resultado pelo respectivo desvio- padrão da variável.

Ainda para verificar se a padronização da variável Idade Média realmente era necessária, aplicou-se a análise de *Cluster* com essa variável em sua escala original e observou-se se essa variável fazia com que cidades com índice de sinistralidade e índice de frequência bem diferentes ficassem juntas em um mesmo *Cluster* somente por ter uma idade média similar, como é o caso das cidades Seropédica e Rio Bonito (ver Tabela 2).

TABELA 2: Teste de para avaliar necessidade de padronização da variável Idade Média

CIDADE	IS	F	Idade Média	IP	IVP	Cluster	Região
SEROPEDICA	0,3198	0,0526	**9**	0,0019	0,0025	5	METROPOLITANA
RIO BONITO	0,0000	0,0000	**9**	0,0005	0,0004	5	METROPOLITANA

A Tabela 2 tem o objetivo de demostrar que os valores das variáveis são bastante diferentes entre as cidades Seropédica e Rio Bonito, exceto a idade média, porém essas cidades ficaram em um mesmo grupo ao gerar a análise de *cluster* sem padronizar a variável idade média.

A Tabela 3 apresenta as estatísticas descritivas dos dados, porém agora com a variável idade média padronizada.

TABELA 3: Estatísticas Descritiva das variáveis

Variável	N	Média	Desvio Padrão	Mínimo	Mediana	Máximo
IS	28	0,7100	0,7710	0,0000	0,6550	3,6570
F	28	0,0521	0,0328	0,0000	0,0569	0,1207
Idade Média	28	2,5755	0,3027	1,8444	2,5328	3,2789
IP	28	0,0355	0,1122	0,0005	0,0033	0,5961
IVP	28	0,0355	0,1095	0,0001	0,0034	0,5801

Ressalta que a variável IS – Índice de Sinistralidade também apresenta valor de desvio padrão elevado, porém como seus valores estão na mesma escala de unidade que as demais variáveis, optou-se em não padronizá-la.

Destaca-se que a padronização da variável Idade Média foi realizada inicialmente sobre a base completa com 68 cidades analisadas, porém a análise estatística foi realizada apenas com 41% (28 cidades) dessas cidades que possuiam volume de dados significativos. Por este motivo, ao observar a Tabela 3 verifica-se que o desvio-padrão da Idade Média Padronizada é 0,3027 e não 1 como esperado. Completa-se ainda que, o desvio padrão da variável idade média das 28 cidades analisadas seria 0,8058.

Com os pontos anteriores avaliados e validados, o passo seguinte é a avaliação do número ideal estimado de *Clusters* utilizado para o estudo.

4 RESULTADO DE PESQUISA

4.1 Determinação do número de grupos

Alguns dos principais métodos de ligação da Análise de *Cluster* foram aplicados sobre as variáveis com intuito de validar a coerência na quantidade de *clusters* formados. A distância Euclidiana foi utilizada para comparação dos grupos em cada passo de agrupamento, exceto para o método Ward no qual se usou a distância Euclidiana ao quadrado. A Tabela 4 a seguir apresenta os resultados encontrados sendo possível verificar que o número de *clusters* varia entre 5 e 6 grupos considerando os métodos testados e usando o coeficiente de correlação intra-classe (R^2). Apenas no método de ligação simples (single linkage) obteve-se um valor de R^2 abaixo de 70%.

TABELA 4: Principais Métodos de Ligação da Análise de Cluster

Linkage	Distancia	g	SST	SSW	SSB	R^2(%)
Average	Euclidiana	06	19, 2187	2,5788	16,6399	86,58
Complete	Euclidiana	06	19,2187	2,1021	17,1166	89,06
Single	Euclidiana	05	19,2187	6,6359	12,5828	65,47
Ward	**Euclidiana ao quadrado**	**05**	**19,2187**	**2,3361**	**16,8826**	**87,84**

Outro teste foi realizado usando os mesmos métodos de ligação da Tabela 4, porém com a medida Euclidiana ao quadrado. O resultado obtido é apresentado na Tabela 5, sendo possível verificar que o número de *Clusters* é reduzido e varia de 2 a 5 com valores baixos de R^2, exceto no método de ligação completo (Complete linkage) que teve resultado similar a Ward. Com estas análises optou-se em estudar o resultado do agrupamento encontrado a partir do método de ligação Ward e distância Euclidiana ao Quadrado já que esse apresentou um bom valor de R^2 e um menor número de grupos comparado aos resultados dos outros métodos.

TABELA 5: Principais Métodos de Ligação da Análise de Cluster – Distância Euclidiana ao quadrado

Linkage	Distancia	G	SST	SSW	SSB	R^2
Average	Euclidiana ao quadrado	3	19,2187	8,2069	11,0118	57,30%
Complete	Euclidiana ao quadrado	5	19,2187	2,7348	16,4839	85,77%
Single	Euclidiana ao quadrado	2	19,2187	9,9939	9,2248	48,00%

Para a análise de determinação da quantidade de grupos foi realizada inicialmente o teste de análise gráfica: análise do salto da diferença de distâncias (ver Figura 4), análise do salto da diferença da similaridade (ver Figura 5) em cada passo do agrupamento. Destaca-se que o cálculo da diferença dessas medidas foi realizado sobre os valores retornados pelo *Minitab*. O terceiro gráfico analisado é o gráfico que apresenta o ponto de inflexão da estatística *Pseudo F* e o coeficiente intra-classe (R^2), vide Figura 6.

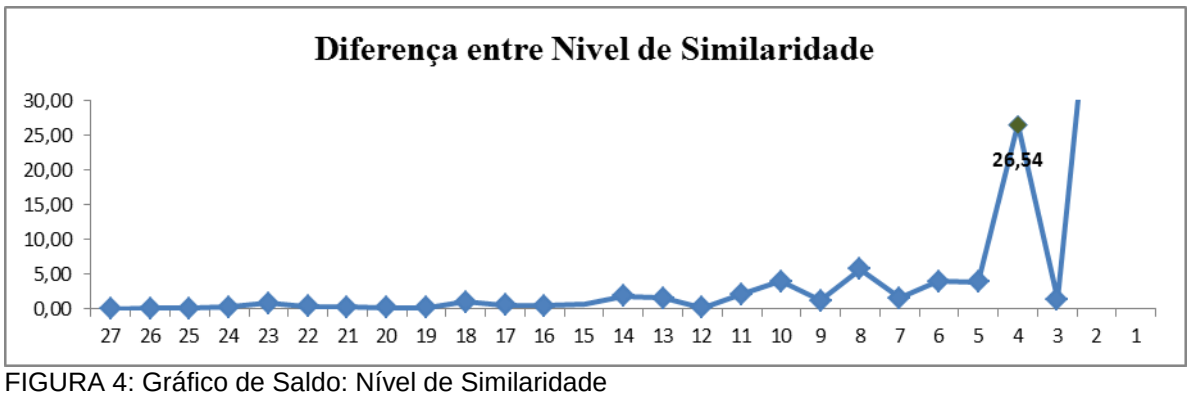


FIGURA 4: Gráfico de Saldo: Nível de Similaridade

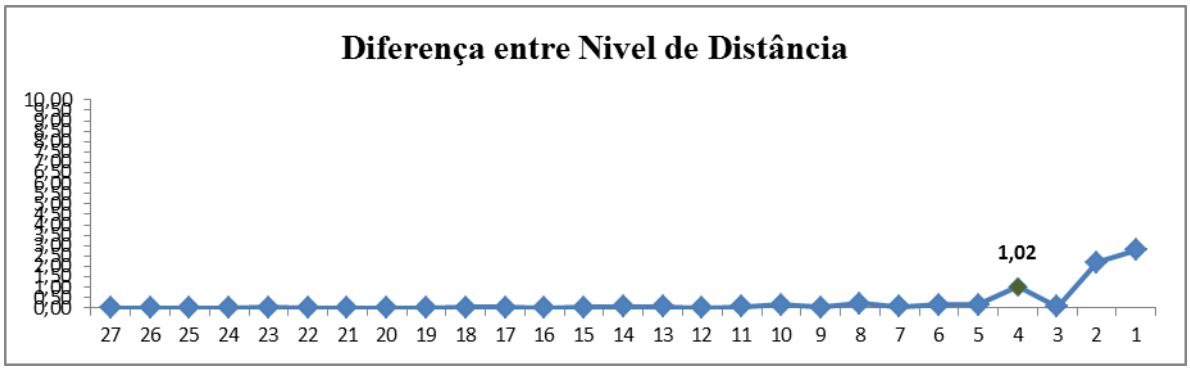


FIGURA 5: Gráfico de Saldo: Nível de Distância

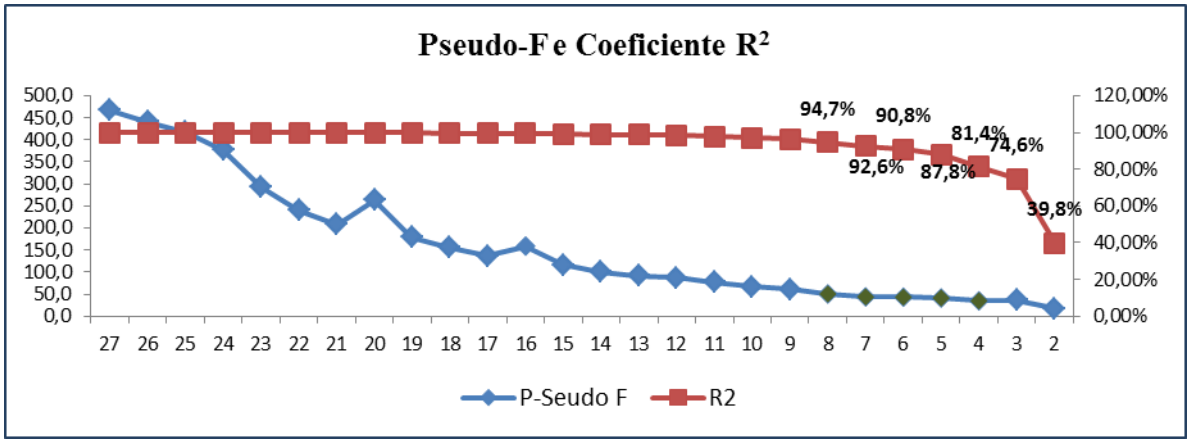


FIGURA 6: Gráfico de Saldo: Estatística F e Coeficiente de Variabilidade Total da partição

A partir dos gráficos das Figuras 5 e 6 verifica-se que o salto ou o aumento significativo das medidas de similaridade e da distância acontece entre os passos relativos a número de grupos 5 e 4. Observando a Figura 07, a partir do número de grupos 8 começa a surgir uma pequena curva onde o ponto de inflexão poderia ser o 4 ou 3.

A Tabela 6 apresenta os resultados obtidos nos passos do agrupamento gerados ao executar o método de agrupamento de ligação Ward usando distância Euclidiana ao quadrado, bem como as diferenças numéricas dos níveis de similaridade e distância apresentadas nas Figuras 05 e 06.

TABELA 6: Análise de Agrupamento – Método Ward

A análise de agrupamento das observações					
Step	**N° Cluster**	**Nível Similaridade**	**Nível Distância**	**Dif_Simil***	**Dif_Dist****
1	27	98,5329	0,056	0,00	0,00
2	26	98,3914	0,062	0,14	0,01
3	25	98,2438	0,067	0,15	0,01
4	24	97,9501	0,079	0,29	0,01
5	23	97,1156	0,111	0,83	0,03
6	22	96,7233	0,126	0,39	0,02
7	21	96,3987	0,138	0,32	0,01
8	20	96,2185	0,145	0,18	0,01
9	19	96,0011	0,153	0,22	0,01
10	18	95,008	0,192	0,99	0,04
11	17	94,476	0,212	0,53	0,02
12	16	94,0003	0,230	0,48	0,02
13	15	93,3663	0,254	0,63	0,02
14	14	91,5009	0,326	1,87	0,07
15	13	89,8795	0,388	1,62	0,06
16	12	89,6748	0,396	0,20	0,01
17	11	87,5752	0,477	2,10	0,08
18	10	83,6105	0,629	3,96	0,15
19	9	82,3535	0,677	1,26	0,05
20	8	76,6469	0,896	5,71	0,22
21	7	75,0718	0,956	1,58	0,06
22	6	71,0628	1,110	4,01	0,15
23	**5**	**67,1652**	**1,260**	**3,90**	**0,15**
24	**4**	**40,6209**	**2,278**	**26,54**	**1,02**
25	3	39,1949	2,333	1,43	0,05
26	2	-18,3798	4,541	57,57	2,21
27	1	-90,714	7,316	72,33	2,77

*Dif_Sim: Diferença de Similaridade

**Dif_Dis: Diferença de Distância

Já na Tabela 7, cujos resultados originaram o gráfico da Figura 7, apresenta-se os cálculos da estatística do Pseudo-F e o coeficiente de correlação intra-classe (R^2) em cada passa (ou seja cada passo possível da partição). Observa que de 5 grupos até 8 a diferença de uma partição para outra nas medidas R^2 e Pseudo F são pequenas, ao passo que de 5 para 4 essa diferença aumenta, sendo mais uma indicação de que trabalhar com uma partição de 5 *clusters* seria adequado.

TABELA 7: Medidas da Estatística F e Coeficiente de Variabilidade Total

g	SST	SSW	SSB	P-Seudo F	R^2
27	19,2187	0,0016	19,2171	466,6	99,99%
26	19,2187	0,0035	19,2152	440,7	99,98%
25	19,2187	0,0058	19,2129	417,1	99,97%
24	19,2187	0,0088	19,2099	377,5	99,95%
23	19,2187	0,0149	19,2038	293,1	99,92%
22	19,2187	0,0228	19,1959	240,6	99,88%
21	19,2187	0,0323	19,1864	207,7	99,83%
20	19,2187	0,0308	19,1879	262,7	99,84%
19	19,2187	0,0537	19,1650	178,4	99,72%
18	19,2187	0,0721	19,1466	156,3	99,63%
17	19,2187	0,0961	19,1226	136,9	99,50%
16	19,2187	0,0961	19,1226	159,3	99,50%
15	19,2187	0,1520	19,0667	116,5	99,21%
14	19,2187	0,2052	19,0135	99,8	98,93%
13	19,2187	0,2579	18,9608	91,9	98,66%
12	19,2187	0,3157	18,9030	87,1	98,36%
11	19,2187	0,4133	18,8054	77,3	97,85%
10	19,2187	0,5571	18,6616	67,0	97,10%
9	19,2187	0,7199	18,4988	61,0	96,25%
8	**19,2187**	**1,0272**	**18,1915**	**50,6**	**94,66%**
7	**19,2187**	**1,4153**	**17,8034**	**44,0**	**92,64%**
6	**19,2187**	**1,7693**	**17,4494**	**43,4**	**90,79%**
5	**19,2187**	**2,3361**	**16,8826**	**41,6**	**87,84%**
4	**19,2187**	**3,5760**	**15,6427**	35,0	**81,39%**
3	19,2187	4,8786	14,3401	36,7	74,62%
2	19,2187	11,5700	7,6487	17,2	39,80%

(*) SSW: Soma de Quadrados dentro dos grupos; SSB: Soma de Quadrados entre grupos; SSTc: soma de quadrados total corrigida.

Com base nas evidências apresentadas até o momento, opta-se em avaliar as principais medidas da *Análise de Cluster*, com intuito de verificar qual a quantidade de grupos ideal para base em estudo (ver Tabela 8).

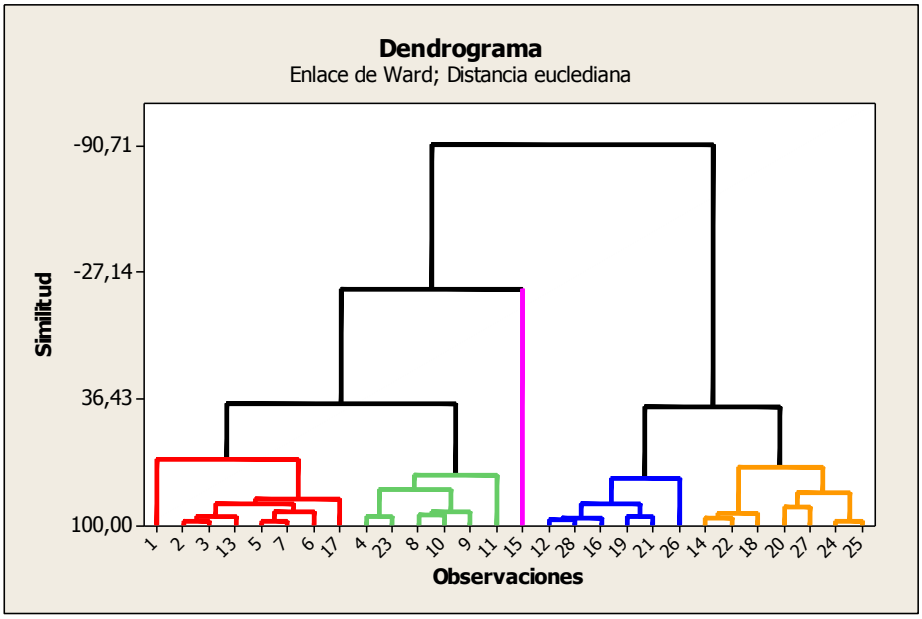
TABELA 8: Medidas da Análise de Cluster

N° Clusters	Similaridade	Distância	R^2 (%)	Diferença Similaridade entre os Passos	Diferença Distância entre os Passos
8	76,647	0,896	94,66	5,707	0,219
7	75,072	0,956	92,64	1,575	0,060
6	71,063	1,110	90,79	4,009	0,154
5	**67,165**	**1,260**	**87,84**	**3,898**	**0,150**
4	40,621	2,278	81,39	26,544	1,018

Observa-se que a medida de similaridade entre os conglomerados decrescem e a distância aumenta quanto menor o número de *clusters*.

O saldo maior acontece entre os números de grupos 5 e 4, a diferença de distância é de 1,02 e a similaridade reduz 26,54, pois altera de 67,02 para 40,6 (ver Tabela 07).

Com a avaliação do dendograma também é possível identificar a separação dos grupos de acordo com a estimativa do número de grupos utilizada.



Fonte: *MINITAB*: Dados de Pesquisa

FIGURA 7: Dendograma

Tanto os gráficos de salto como o dendograma consideram a distância entre os grupos formados. O número sugerido de *clusters* varia de 8 a 5, embora a quantidade de 5 *clusters* aparente ser melhor.

Adotou-se o método de agrupamento das K-Médias como um critério de validação da partição escolhida com 5 *Clusters* para comparar se os grupos formados seriam diferentes. Conforme Tabela 9, este critério foi realizado e encontrado as mesmas cidades em cada *Cluster* formado, o que considera que método Ward com 5 *Cluster* seria ideal para os dados em estudo. As sementes de inicialização do método K-Médias foram os vetores de médias obtidos na análise de *clusters* pelo método de Ward para 5 grupos.

TABELA 9: Método K-Médias e Ward (5 grupos)

Cidades	IS	F	Idade_padronizada	IP	IVP	Cluster K-Media	Cluster Ward (5)
RIO DE JANEIRO	0,7767	0,0644	7,124989591	0,5961	0,5801	1	1
NOVA IGUACU	0,8294	0,0664	6,588677065	0,0859	0,0887	1	1
SAO JOAO DE MERITI	0,8406	0,0671	6,76013986	0,0710	0,0824	1	1
BELFORD ROXO	0,7049	0,0715	6,524966262	0,0368	0,0407	1	1
NILOPOLIS	0,6612	0,0570	7,03898051	0,0331	0,0342	1	1
MESQUITA	0,6492	0,0528	6,568	0,0310	0,0328	1	1
MAGE	0,8972	0,0595	6,630952381	0,0042	0,0044	1	1
JAPERI	0,6735	0,0816	5,836734694	0,0024	0,0022	1	1
DUQUE DE CAXIAS	1,3756	0,0896	6,46857671	0,0537	0,0546	2	2
SAO GONCALO	1,0495	0,0569	7,104587156	0,0271	0,0271	2	2
NITEROI	1,2440	0,0812	7,166666667	0,0116	0,0089	2	2
QUEIMADOS	1,0608	0,0878	6,608108108	0,0073	0,0070	2	2
ITAGUAI	1,9078	0,0813	7,146341463	0,0061	0,0063	2	2
ANGRA DOS REIS	1,4813	0,0500	6,4	0,0010	0,0006	2	2
RIO DAS OSTRAS	0,1938	0,0319	6,755319149	0,0047	0,0044	3	3
MARICA	0,0902	0,0182	6,654545455	0,0027	0,0028	3	3
MACAE	0,0000	0,0000	6,12195122	0,0020	0,0015	3	3
SAO PEDRO DA ALDEIA	0,1363	0,0357	6,214285714	0,0014	0,0014	3	3
PIABETA	0,0000	0,0000	4,909090909	0,0005	0,0006	3	3
VILA MURIQUI	0,1433	0,0909	6,727272727	0,0005	0,0001	3	3
ITABORAI	0,2947	0,0263	7,144736842	0,0038	0,0039	4	4
MANGARATIBA	0,4802	0,0698	7,581395349	0,0021	0,0014	4	4
SEROPEDICA	0,3198	0,0526	8,631578947	0,0019	0,0025	4	4
CAMPOS DOS GOYTACAZES	0,4004	0,0455	7,318181818	0,0011	0,0005	4	4
PARACAMBI	0,0000	0,0000	7,842105263	0,0009	0,0012	4	4
ARARUAMA	0,0000	0,0000	7,692307692	0,0006	0,0006	4	4
RIO BONITO	0,0000	0,0000	8,727272727	0,0005	0,0004	4	4
CABO FRIO	3,6566	0,1207	5,655172414	0,0029	0,0014	5	5

4.2 Análise dos Resultados Obtidos

Dentre as cidades utilizadas no agrupamento, 65% são da região metropolitana do Rio de Janeiro, o que já era esperado, visto se tratar da região de maior massa da carteira.

Opta-se em analisar a partição com 5 *clusters*, visto que nas análises apresentadas na seção 4.1 foi o número de maior evidência de que poderia resultar na melhor partição. A partição obtida é apresentada na Tabela 10.

TABELA 10: Partição com 5 Clusters

Cluster/Cidade	Região	Média				
1		IS	F	Idade	IP	IVP
BELFORD ROXO	METROPOLITANA	75%	7%	2	11%	11%
JAPERI	METROPOLITANA					
MAGE	METROPOLITANA					
MESQUITA	METROPOLITANA					
NILOPOLIS	METROPOLITANA					
NOVA IGUACU	METROPOLITANA					
RIO DE JANEIRO	METROPOLITANA					
SAO JOAO DE MERITI	METROPOLITANA					
2						
ANGRA DOS REIS	BAIA DA ILHA GRANDE	135%	7%	3	2%	2%
DUQUE DE CAXIAS	METROPOLITANA					
ITAGUAI	METROPOLITANA					
NITEROI	METROPOLITANA					
QUEIMADOS	METROPOLITANA					
SAO GONCALO	METROPOLITANA					
3						
MACAE	NORTE FLUMINENSE	9%	3%	3	0,2%	0,2%
MARICA	METROPOLITANA					
PIABETA	SERRANA					
RIO DAS OSTRAS	BAIXADA LITORANEA					
SAO PEDRO DA ALDEIA	BAIXADA LITORANEA					
VILA MURIQUI	BAIA DA ILHA GRANDE					
4						
ARARUAMA	BAIXADA LITORANEA	21%	3%	3	0,2%	0,2%
CAMPOS DOS GOYTACAZES	NORTE FLUMINENSE					
ITABORAI	METROPOLITANA					
MANGARATIBA	BAIA DA ILHA GRANDE					
PARACAMBI	METROPOLITANA					
RIO BONITO	METROPOLITANA					
SEROPEDICA	METROPOLITANA					
5						
CABO FRIO	BAIXADA LITORANEA	366%	12%	2	0,3%	0,1%

(*) idade média padronizada

Antes de avaliar as estatísticas descritivas de todos os *clusters* formados na partição, fez-se análise dos *clusters* avaliando a dispersão das regiões dentre os grupos. Em todos os clusters formados há cidades de diferentes regiões, exceto o cluster de número 1 que contém todas as cidades da região metropolitana do Rio de Janeiro.

Foi obtido um cluster que pode ser dado como exceção, que é o Cluster 5, em que possui somente uma (1) cidade dentro dele, a cidade de Cabo Frio que de fato é uma cidade com uma característica diferente das outras no que se refere a variável índice de sinistralidade (IS).

Mesmo se a partição fosse de 4 *Clusters* a cidade de Cabo Frio não se agruparia com outras visto que a severidade e quantidade de sinistro registrada nesta cidade é muito em alta em relação as demais cidades do Rio de Janeiro, com isso induz ao índice de sinistralidade e a frequência serem maiores.

Ao comparar o agrupamento realizado pelas Seguradoras, por proximidade geográfica, apenas o *Cluster* 1 é semelhante visto que todas as cidades do grupo pertencem à mesma região, Metropolitana, enquanto que, os demais *Clusters* formados apresentam cidades em um mesmo grupo de regiões diferentes. Como por exemplo, o *Cluster* 3 que contem 6 cidades diferentes correspondentes a 5 diferentes regiões. Geograficamente são distantes entre si, porém pela análise de *Cluster* são cidades com medidas semelhantes, o que mais uma vez comprovaria que talvez seja necessário um estudo aplicando outros pontos para taxação quanto à localidade do veículo.

A Tabela 11 contém as estatísticas descritivas de cada variável para cada grupo formado na partição de 5 grupos pelo método Ward. Desta forma será possível identificar as características em cada agrupamento. Cada *cluster* é formado com a quantidade de 6 a 8 cidades, exceto o *Cluster* de número 5. Este último *Cluster*, por exemplo, é formado apenas com a cidade Cabo Frio da região da Baixada Litorânea, visto conter o índice de sinistralidade e o índice de frequência de sinistro muito acima da média dos demais *clusters*.

O *Cluster* 1, contendo apenas cidades da região metropolitana, cujo índice de sinistralidade varia entre 64,92% a 89,72%, frequência de sinistro entre 5% e 8,16%. Todas as cidades desse grupo com índice de produção e proporção de veículos populares significativos. A variável idade média padronizada é a de menor influência no agrupamento, visto que esta variando entre 2 e 3 anos.

O *Cluster* 2 se destaca pelo o índice de sinistralidade das cidades acima de 100%. Embora o índice de frequência seja próximo do *Cluster* 1, as demais variáveis se diferenciam em seus valores como é caso do índice de produção e índice de veículo popular que esta em média 1,78% e 1,74% respectivamente.

Já os *Clusters* 3 e 4 são formados por cidades pertencentes a regiões diferentes, porém com semelhança em todas as variáveis, exceto no índice de sinistralidade que no *Cluster* 03 vai até 19,38% e o *Cluster* 04 tem sinistralidade máxima de 48,02%.

TABELA 11: Estatística descritiva dos grupos formados na partição com 5 clusters

Variável	cluster	N	Média	Desvio Padrão.	Mínimo	Mediana	Máximo
IS	1	8	0,7541	0,0946	0,6492	0,7408	0,8972
	2	6	1,3530	0,3210	1,0490	1,3100	1,9080
	3	6	0,0939	0,0798	0,0000	0,1132	0,1938
	4	7	0,2136	0,2084	0,0000	0,2947	0,4802
	5	1	3,6566	*	3,6566	3,6566	3,6566
Freq	1	8	0,0651	0,0090	0,0528	0,0654	0,0816
	2	6	0,0745	0,0168	0,0500	0,0813	0,0897
	3	6	0,0295	0,0337	0,0000	0,0250	0,0909
	4	7	0,0277	0,0289	0,0000	0,0263	0,0698
	5	1	0,1207	*	0,1207	0,1207	0,1207
Idade Média Padronizada	1	8	2	0,1472	2	2	3
	2	6	3	0,1357	2	3	3
	3	6	2	0,2640	2	2	3
	4	7	3	0,2305	3	3	3
	5	1	2	*	2	2	2
Ind_Prod	1	8	0,1076	0,1995	0,0024	0,0349	0,5961
	2	6	0,0178	0,0197	0,0010	0,0095	0,0537
	3	6	0,0020	0,0016	0,0005	0,0017	0,0047
	4	7	0,0016	0,0011	0,0005	0,0011	0,0038
	5	1	0,0029	*	0,0029	0,0029	0,0029
Ind_Pop	1	8	0,1082	0,1933	0,0022	0,0375	0,5801
	2	6	0,0174	0,0203	0,0007	0,0079	0,0546
	3	6	0,0018	0,0016	0,0001	0,0015	0,0044
	4	7	0,0015	0,0013	0,0004	0,0012	0,0039
	5	1	0,0014	*	0,0014	0,0014	0,0014

*O Cluster 5 é formado apenas por uma cidade, por este motivo não tem valor de desvio-padrão.

Através do gráfico apresentado na Figura 8 é possível entender melhor a análise em síntese da Tabela 11.

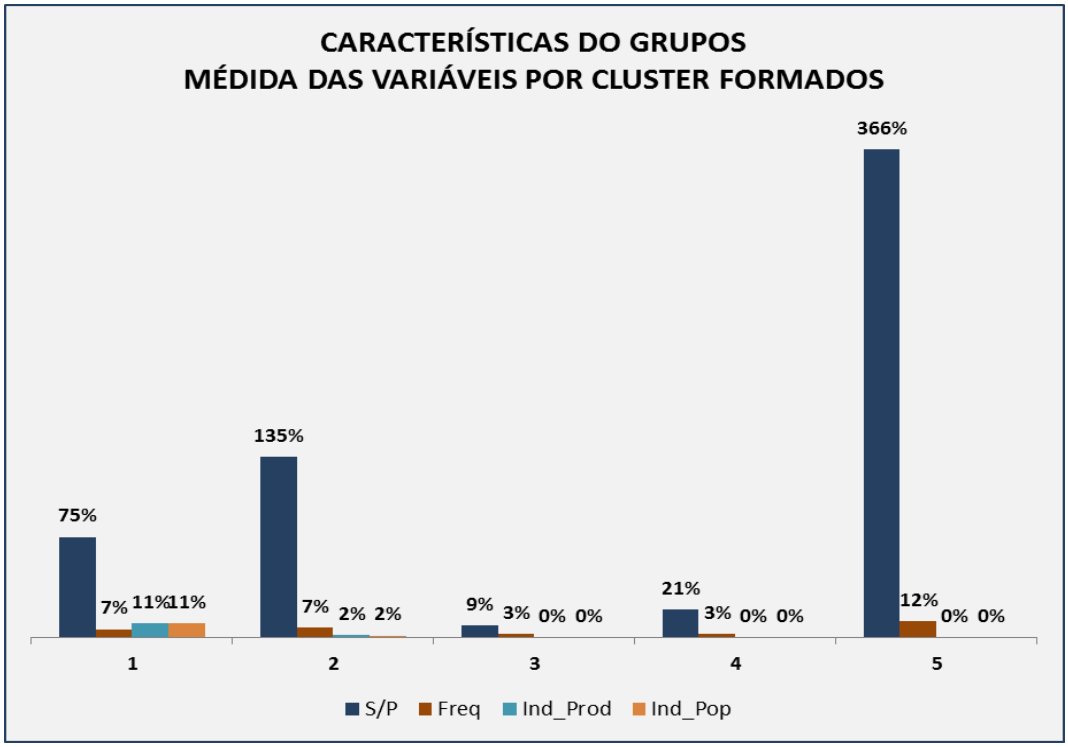


FIGURA 8: Gráfico – Características dos Grupos

Verifica-se, na Figura 8, que todos os grupos apresentam uma distância entre as médias, sugerindo uma boa aplicação do modelo.

Observa-se, ainda, que o *Cluster* 1 concentra as cidades que obtiveram os melhores resultado de sinistralidade (IS) juntamente com volume significativo de itens (ind_prod) e que o *Cluster* 2 reúne as cidades com as maiores sinistralidade (IS). Abaixo apenas do grupo 5 que é formado apenas por uma cidade, Cabo Frio.

Os *Clusters* 3 e 4 apresentam menores valores de sinistralidade e frequência, contudo são formadas por cidades com baixa volume de itens (ind_produção).

5 CONCLUSÃO

O método de precificação bem aplicado em uma seguradora e em um Programação de Proteção Veicular é fundamental para o seu resultado satisfatório, pois o preço fornecido por ela deverá ser suficiente para cobrir as despesas de sinistro.

A pesquisa proposta tem o objetivo de verificar a diferença no agrupamento das cidades do estado do Rio de Janeiro ao aplicar o método estatístico de *Análise de Cluster* sobre os dados de uma associação de Proteção Veicular em comparação com o agrupamento por proximidade geográfica, que é adotado por algumas Seguradoras.

Em resposta a problema proposto na pesquisa, pode-se avaliar que as variáveis, quantidade e valor de sinistros, quantidade de itens exposto e prêmio são fatores decisivos para determinação de regiões.

Com avaliação dos grupos sugeridos na análise de *clusters* apresentada nessa monografia podemos verificar que nas cidades onde se obtém o menor índice de sinistralidade a sua precificação seja suficiente para cobrir despesa de sinistro em comparação às demais e que as cidades com alto índice de sinistralidade e frequência deveriam ter precificações mais elevadas, para que a associação tenha condições de cobrir os sinistros indenizados naquela região.

Para as demais regiões o prêmio deve ser variável juntamente com os sinistros daquela região.

Partindo deste pressuposto nota-se uma possível necessidade de segmentar as regiões do estado do Rio de Janeiro através de algo mais seletivo, do que somente regionalização como utilizado pela seguradora. Desta forma, foi utilizado o método de *Cluster* para segmentar as cidades do Rio de Janeiro, considerando as variáveis citadas, assim como comparar os agrupamentos obtidos, com os agrupamentos utilizados pela associação.

Pode ser notada uma diferença entre os dois tipos de agrupamentos, principalmente no agrupamento geográfico, pois desta forma cidades que possuem variados índice de sinistro estão agrupados nas mesmas regiões.

Regiões com altos índices de sinistros deveriam ser taxados de forma parecida, e serem tratados de forma diferentes das regiões que possuem um índice de sinistro menor.

Conforme apresentado existem algumas cidades que deveriam ser tratadas de forma especial devido às características que essas possuem.

Cidades com grande volume de sinistro como Cabo Frio, no agrupamento realizado pela associação, elevando o índice de sinistralidade de todo o Estado do Rio de Janeiro e possivelmente contribuindo para seleção adversa do risco, pois com taxa atribuída na proporção indesejável os riscos considerados bons não entrariam na base ou apenas os riscos agravados entram elevando ainda mais o índice de sinistralidade.

Portanto, comprova-se que existe uma diferença entre o agrupamento por região, e o agrupamento por *cluster*, considerando variáveis relacionadas a sinistro e prêmio.

REFERÊNCIAS BIBLIOGRÁFICAS

AAKER, D. A.; KUMAR, V.; DAY, G. S. **Pesquisa de marketing**, São Paulo: Atlas, 2001. 745p.

ABRANTES Jose. **Associativismo e Cooperativismo**. São Paulo: Interciência, 2004.

ANDERBERG, M. R. **Cluster analysis for applications**. New York: Acafenic press, 1973, 359p

ANDERSON, T. W. **An introduction to multivariate statistical analysis**, New York: John Wiley & Sons, 1984, 675 p.

BARROSO, L. P., ARTES, **R. Análise de Multivariada**. Lavras: UFLA, 2003. 157p.

BRASIL, Gilberto. **O ABC da Matemática Atuarial e Princípios Gerais de Seguro.** Pôrto Alegre: Sulina 1985.

BUSSAB, W. DE O; MIAZAKI, E. S; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 105p.

CORMACK, R. **A review of classification. Journal of the Royal Statistical Society** (Series A), v.134, p.321-367, 1971.

CORRAR, Luiz J.; PAULO, Edílson; DIAS FILHO, José Maria (Coords.). **Análise multivariada: para os cursos de administração, ciências contábeis e economia**. São Paulo: Atlas, 2007. 541 p.

Fundação Escola Nacional de Seguros (Brasil). **Teoria geral do seguro**. 3. ed. Rio de Janeiro: FUNENSEG, 2001. 110p.

FUNENSEG, Escola Nacional de Seguros. **Teoria Geral do Seguro**. 7. ed. Rio de Janeiro: Funenseg, 1996.62p.

FUNENSEG. DIRETORIA DE ENSINO E PESQUISA. **Teoria geral do seguro I**. Acessoria tecnica de Jose Antonio Menezes Varanda; Ordenacao didatica de Marilia Scofano de Souza Aguiar. 3. ed. Rio de Janeiro:Funenseg, 2005.

GASPARINI Diogenes. **Direito Administrativo**. 15 ed. São Paulo: Saraiva, 2010.

RANDALL, Everett. Introdução à Subscrição. Rio de Janeiro: Funenseg, 2000. 216p.

MANO, Cristina Maria Cantanhede Amarante Biasotto. **Melhoria da qualidade na tarifação de seguros** : uso de modelos de credibilidade. Rio de Janeiro: FUNENSEG, 1997. 103p.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada**: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

REIS, E.; **Estatística multivariada aplicada**. Lisboa: Edições Silabo, 1997. 342p.

SOUZA, Silney de. **Seguros: contabilidade, atuária e auditoria**. 2. ed. rev. e atual. São Paulo: Saraiva, 2007. xvii, 229p.

WARD, J. H.; **Hierarchical grouping to optimize an objective function.** Journal of. American Statistical Association, v. 58, p. 236-244, 1963.

TUDO SOBRE SEGUROS. **Fatos e Indicadores do Mercado.**
Disponível em: <http://www.tudosobreseguros.org.br/sws/portal/pagina.php?l=267>
Acessado em 17/01/2013

TUDO SOBRE SEGUROS. **Entenda o Seguro Automóvel.**
Disponível em <http://www.tudosobreseguros.org.br/sws/portal/pagina.php?l=167#o que e franquia> Acessado em 20/02/2013

SINCOR-MG. **Reflexão sobre associações de seguros e o programa de proteção automotiva**. Disponível em < http://revistaapolice.com.br/2012/09/reflexoes-sobre-associacoes-de-seguros-e-o-programa-de-protecao-automotiva/> Acessado 31/10/2012

JORNAL DO COMERCIO. **A importância das Cooperativas.**
Disponível em <http://www.cooperativismo.org.br/cooperativismo/noticias/noticia.asp?id=19277>
Acessado em 24/10/2012

Everitt, B.S., Landau, S, Leese, M. Cluster Analysis.New York: Oxford University Press, 2001, 237 p.

CALINSKI, T.; Harabasz, J. A dendrite method for cluster analysis. **Communications in Statistics,** Londres, v.3, p.1-27, 1974.