UNIVERDADE FEDERAL DE MINAS GERAIS

Especialização em Estatística com Ênfase em Indústria e Mercado

ANÁLISE DE AGRUPAMENTO DE OPERADORAS DE PLANOS DE ASSISTÊNCIA À SAÚDE:

Estudo de caso da carteira de clientes de uma consultoria atuarial

Geisa Andressa Corrêa da Silva

Belo Horizonte

2015

Geisa Andressa Corrêa da Silva

ANÁLISE DE AGRUPAMENTO DE OPERADORAS DE PLANOS DE ASSISTÊNCIA À SAÚDE:

Estudo de caso da carteira de clientes de uma consultoria atuarial

Monografia apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais, como parte dos requisitos para obtenção do título de Especialista em Estatística

Área de Concentração: **Ênfase em Indústria e Mercado**

Orientadora: Profa. Sueli Aparecida Mingoti – Ph. D. em Estatística

Belo Horizonte

2015

AGRADECIMENTOS

Agradeço a Deus, pela força que me permitiu chegar até aqui.

A professora Sueli Aparecida Mingoti, pela orientação, paciência e atenção durante a realização deste trabalho.

Aos demais professores que mesmo diante das mais diversas dificuldades ajudaram a desenvolver nosso potencial profissional e social.

Aos amigos e colegas de curso pelo incentivo e apoio constante.

A minha família que com muito carinho e compreensão, não mediram esforços para que eu chegasse até esta etapa de minha vida.

LISTA DE ILUSTRAÇÃO

Figura 1 - Histograma das Variáveis 21
Figura 2 - Boxplot das Variáveis 22
Figura 3 - Histograma das Variáveis Padronizadas 23
Figura 4 - Boxplot das Variáveis Padronizadas 24
Figura 5 - Dendograma: agrupamento das operadoras 24
Figura 6 - Escore das duas componentes principais, Ward (k=6) 33
Figura 7- Escore das duas componentes principais, K-médias (k=6) 34
Figura 8 - Boxplot das variáveis padronizadas e agrupadas, Ward k=6 34
Figura 9 - Boxplot das variáveis padronizadas e agrupadas, K-médias k=6 36
Figura 10 - Escore das duas componentes principais, Ward (k=7) 38
Figura 11 - Escore das duas componentes principais, K-médias (k=7) 39
Figura 12 - Boxplot das variáveis padronizadas e agrupadas, Ward k=7 39
Figura 13- Boxplot das variáveis padronizadas e agrupadas, K-médias k=7 41

LISTA DE QUADROS

Quadro 1 - Descrição das variáveis utilizadas no estudo 19
Quadro 2 - Estatísticas descritivas das variáveis (*n*=33 operadoras) 20
Quadro 3 - Estatísticas descritivas das variáveis padronizadas (*n*=33 operadoras) 22
Quadro 4 - Histórico do agrupamento das 33 operadoras - Método Ward 25
Quadro 5 - Soma de quadrados das Partições, R^2 e Pseudo *F* 26
Quadro 6 - Agrupamento Ward - k=3 27
Quadro 7 - Agrupamento Ward - k=4 27
Quadro 8 - Agrupamento Ward - k=5 27
Quadro 9 - Agrupamento Ward - k=6 28
Quadro 10 - Agrupamento Ward - k=7 28
Quadro 11- Agrupamento K-Médias - k=3 30
Quadro 12 - Agrupamento K-Médias - k=4 30
Quadro 13- Agrupamento K-Médias - k=5 30
Quadro 14 - Agrupamento K-Médias - k=6 31
Quadro 15 - Agrupamento K-Médias - k=7 31
Quadro 16 - Resultados por método de análise de *cluster* 32
Quadro 17 - Estatística Descritiva das variáveis padronizadas, Ward k=6 35
Quadro 18 - Estatística Descritiva das variáveis padronizadas, K-Médias k=6 37
Quadro 19 - Estatística Descritiva das variáveis padronizadas, Ward k=7 40
Quadro 20 - Estatística Descritiva das variáveis padronizadas, K-médias k=7 42

RESUMO

Este trabalho trata da aplicação da técnica estatística de análise de agrupamento como instrumento para agregar operadoras de planos de saúde, através da utilização de quatro variáveis relevantes no mercado de saúde suplementar: despesa assistencial, receita de contraprestação, despesa administrativa e beneficiários. O mercado de saúde suplementar sofreu várias mudanças nos últimos anos, muitas delas vindas da regulação do setor pela Agência Nacional de Saúde Suplementar. É papel da consultoria atuarial especializada em saúde trabalhar no desenvolvimento de metodologias que ajudem a gerir um mercado com tantos riscos.

Palavras-chaves: Saúde Suplementar, Técnicas de Estatística Multivariada, Análise de Agrupamento, Operadora de Planos de Saúde

SUMÁRIO

1 INTRODUÇÃO 6
1.1 Objetivo Geral 6
1.1.1 Objetivos específicos 7
2 FUNDAMENTAÇÃO TEÓRICA 7
2.1 Saúde Suplementar 7
2.2 Análise Estatística Multivariada 8
2.3 Análise de Agrupamento 9
2.3.1 Método Ward 13
2.3.2 Estimação do Número de Clusters da Partição 15
2.3.3 Método das K-Médias 16
3 Descrição das Variáveis e Banco de Dados 18
4 Análise Estatística dos Dados 19
4.1 Análise de agrupamento pelo método Ward 24
4.2 Análise de agrupamento pelo método das K-médias 28
5 CONSIDERAÇÕES FINAIS 43
REFERÊNCIAS 45
ANEXO I 47

1 INTRODUÇÃO

O atendimento privado de saúde pode ser definido como saúde suplementar, onde empresas denominadas operadoras de planos de saúde ofertam este serviço. Este trabalho estuda através da utilização da técnica estatística multivariada de análise de agrupamento, quatro variáveis observadas em trinta e três operadoras de planos de saúde que contemplam uma carteira de uma consultoria atuarial do mercado de saúde suplementar brasileiro.

Os custos com a saúde aumentam a cada ano e inúmeros são os fatores que tornam este um mercado de risco e complexo, deixando muitas dessas empresas operando com dificuldades financeiras. Pode-se citar como desafios envelhecimento da população que ocasiona uma maior utilização do plano por uma faixa etária que gera maiores custos, a evolução das tecnologias que surgem na área de saúde com equipamentos e remédios cada vez mais caros, o aumento periódico dos procedimentos mínimos a serem cobertos pelo plano, e existem outras inúmeras situações vividas pelas operadoras que influenciam em aumentos cada vez mais expressivos das suas despesas.

A Atuária como ciência que estuda riscos financeiros e econômicos, atua neste mercado inicialmente com o objetivo de estimar custos que gere planos acessíveis para o mercado em que a operadora se encontra e suficientes para arcar com a utilização dos serviços médicos/hospitalares pelos beneficiários e os gastos com a administração do negócio.

Em um cenário tão conturbado é papel da consultoria especializada em saúde auxiliar as operadoras em sua gestão. A cada dia tem-se mais evidências de que é necessário rever os modelos utilizados e aplicar diferentes técnicas para atender as necessidades impostas pelo mercado.

1.1 Objetivo Geral

O objetivo deste estudo é utilizar a técnica de análise de *cluster* para criar grupos de operadoras de planos de saúde que possuam características similares entre si. Através desta análise propõe-se classificar as operadoras a partir das suas características

principais. Este trabalho permitirá à consultoria atuarial direcionar melhor sua mão-de-obra no relacionamento com seus clientes, verificar as principais deficiências e criar planos de melhoria para cada grupo formado, tornando seu atendimento mais personalizado. Essas decisões irão permitir que a consultoria atuarial melhore a gestão da sua carteira de clientes.

1.1.1 Objetivos específicos

Com o intuito de alcançar o objetivo geral, foram adotados os seguintes procedimentos:

- Utilização das técnicas de Análise de *Cluster* de Ward e K- Médias para agrupar as operadoras através das variáveis selecionadas;
- Avaliação dos resultados obtidos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Saúde Suplementar

Segundo Cordeiro (1984) o mercado de saúde suplementar começou a se estruturar na década de 60, iniciando nesta época os primeiros contratos médicos em São Paulo e no Rio de Janeiro. Desde então, o setor vem se desenvolvendo. De acordo com Almeida (1998), 28% da população urbana brasileira é coberta pela saúde suplementar e mais de 20% dos gastos com saúde é proveniente deste setor.

Com o advento da regulação através da criação da Lei 9.656/98 com o objetivo de “disciplinar o mercado de prestação de serviço suplementar à saúde em todo país”, (Brasil, 1999), e no ano seguinte a criação da Agência Nacional de Saúde Suplementar (ANS) como agência reguladora do setor, houve significativas mudanças no mercado.

A primeira atuação foi econômico-financeira com a instituição de garantias financeiras obrigatórias para as operadoras atuarem no mercado e em seguida pela

preocupação assistencial, uma série de normativas que visavam melhorar a qualidade da prestação de serviço aos beneficiários (BRASIL, 2011).

No Brasil, são seis as modalidades de compra e venda de serviços médicos: Medicina de Grupo, Cooperativas Médicas, Seguros Saúde, Autogestão, Filantropia e Administradoras de Benefícios.

As administradoras de benefícios são empresas que em caráter de estipulante ou na prestação de serviço a outra pessoa jurídica lida com a contratação de planos de saúde. A autogestão é uma modalidade dirigida a um grupo fechado de pessoas que pertençam a mesma classe profissional ou tenham vínculo com a empresa que irá gerir o plano. As cooperativas médicas são empresas que na forma de regime jurídico de cooperativa operam planos de assistência à saúde. As filantropias médicas são pessoas jurídicas sem fins lucrativos que operam planos de saúde. As seguradoras especializadas em saúde são as sociedades securitárias que trabalham ofertando planos. A medicina de grupo são as operadoras restantes que não se enquadram nos grupos anteriormente especificados (BRASIL, 2014).

As operadoras presentes neste estudo são das quatro principais modalidades presentes no mercado: Medicina de Grupo, Cooperativa Médica, Filantropia e Autogestão.

2.2 Análise Estatística Multivariada

A utilização de técnicas estatísticas para análise e avaliação de dados como uma ferramenta prática para o mercado é cada vez mais comum na tomada de decisões gerenciais. De acordo com Hair et al. (2005, p. 25) “técnicas analíticas multivariadas estão sendo amplamente aplicadas na indústria, no governo e em centros de pesquisa acadêmica.”

A análise multivariada, segundo Mingoti (2005), trata de um conjunto de métodos estatísticos que permite a análise simultânea de medidas múltiplas para cada indivíduo ou objeto em análise, ou seja, qualquer método estatístico que permita a análise simultânea de duas ou mais variáveis pode ser considerado como multivariado.

Corrar et al. (2014) complementam que:

> "A análise multivariada pode ser definida como o conjunto de métodos que permitem a análise simultânea dos dados recolhidos para um ou mais conjuntos de indivíduos (populações ou amostras) caracterizados por mais de duas variáveis correlacionadas entre si, sendo que as variáveis podem ser quantitativas (discretas ou contínuas) ou qualitativas (ordinais ou nominais). Somente as técnicas de estatística multivariada permitem que se explore a performance conjunta das variáveis e se determine a influência ou importância de cada uma, estando as demais presentes." (Corrar et al., 2014 p. 3)

Hair et al. (2005) afirmam que grandes transformações ocorrerão no futuro no modo com que os profissionais de pesquisa analisam seus problemas pela utilização dos métodos de análise multivariada para a resolução de cenários complexos.

2.3 Análise de Agrupamento

A análise de agrupamentos também chamada de análise de conglomerados ou análise de *cluster* "é uma técnica analítica para construir subgrupos de indivíduos ou objetos." (Hair et al., p. 33, 2005).

Segundo Jonhson e Wichern (2007) a análise de agrupamento é uma das técnicas mais primitivas, onde são realizadas suposições sobre quantidade de grupos e estruturas, com base na semelhança ou diferenças nos dados.

"Cluster analysis is a more primitive technique in that no assumptions are made concerning to the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities)."(JONHSON; WICHERN, pag. 671, 2007).

De acordo com Trion e Bailey (1970) o método de *cluster* é um procedimento de estatística multivariada que trabalha com um conjunto de dados contendo informações de uma amostra reorganizando esses dados para que fiquem agrupados de forma homogênea. Pode-se citar como quatro principais objetivos: desenvolvimento de uma tipologia ou classificação, investigação de esquemas conceituais úteis para o agrupamento, geração de hipóteses através da exploração de dados e teste de hipóteses

ou a tentativa de determinar se tipos definidos através de outros procedimentos são de fato presente em um conjunto de dados.

> "A clustering method is a multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups. Most of the varied uses of cluster analysis can be subsumed under four principal goals:
>
> - Development of a typology or classification,
> - Investigation of useful conceptual schemes for grouping entities,
> - Hypothesis generation through data exploration, and
> - Hypothesis testing or the attempt to determine if types defined through other procedures are in fact present in a data set." (Trion e Bailey, p. 23, 1970)

Corrar et al. (2014) citam que os objetivos da análise de conglomerados são a descrição taxonômica pelo seu propósito exploratório e classificação de objetos com base empírica, a simplificação de dados e identificação das relações sendo possível obter uma visão das informações existentes entre as observações.

Segundo Jonhson e Wichern (2007) o objetivo básico da análise de conglomerados é descobrir os agrupamentos naturais dos itens ou variáveis em estudo, mas é necessário primeiro desenvolver uma escala quantitativa para medir a semelhança entre os objetos.

"To summarize, the basic objective in cluster analysis is to discover natural groupings of the items (or variables). In turn, we must first develop a quantitative scale on which to measure the association (similarity) between objects." (JONHSON e WICHERN, pag. 671, 2007).

Hair et al. (2005) afirmam sobre o conceito de similaridade:

> "O conceito de similaridade é fundamental para a análise de conglomerados. A similaridade entre objetos (interobjectsimilarity) é uma medida de correspondência, ou semelhança, entre objetos a serem agrupados." (Corrar, Paulo e Filho, 2014 p. 3)

Segundo Trion e Bailey (1970) apesar do conceito de similaridade ser simples, os procedimentos utilizados para medi-la estão longe de serem simples.

> "The things are recognized as similar or dissimilar is fundamental to the process of classification. Despite its apparent simplicity, the concept of similarity, and especially the procedures used to measure similarity, are far from simple." (Trion e Bailey, p. 25, 1970)

Mingoti (2005) expõe que na utilização da técnica é necessário decidir qual será a medida de similaridade a ser empregada. Tem-se como principais medidas:

- Distância Euclidiana;
- Distância generalizada ou ponderada;
- Distância de Minkowsky;
- Coeficiente de concordância simples;
- Coeficiente de concordância positiva;
- Coeficiente de concordância de Jaccard;
- Distância Euclidiana média, dentre outras.

De acordo com Jonhson e Wichern (2007) dificilmente, mesmo com um computador potente é viável examinar todas as possibilidades de agrupamento. Devido a isso tem surgido uma grande variedade de algoritmos para facilitar esta análise.

Segundo Mingoti (2005) o conjunto de técnicas para construção de conglomerados abrangem as hierárquicas nas quais os dados são aglomerados ou divididos passo a passo, e as não-hierárquicas cuja a quantidade de grupos nos quais os elementos podem ser alocados já é pré-definida.

Jonhson e Wichern (2007) afirmam que a técnica de agrupamento hierárquica é uma série de fusões ou divisões sucessivas. Quando ela é aglomerativa cada objeto é um grupo e a cada passo é formado um novo grupo, os objetos mais semelhantes são agrupados primeiro, e esses grupos são mesclados de acordo com suas semelhanças até que todos os subgrupos são fundidos em um único *cluster*.

> "Hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions. Agglomerative hierarchical methods start with the individual objects. Thus, there are initially as many clusters as objects. The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster." (JONHSON; WICHERN, pag. 671, 2007).

Neste método apenas dois grupos podem ser unidos em cada passo do algoritmo de agrupamento (Corrar et al.,2014).

Já para o processo divisivo, Jonhson e Wichern (2007), afirmam que este trabalha na direção oposta do aglomerativo. O método se inicia com um único grupo com todos os objetos, que são divididos em dois grupos, e esses subgrupos são divididos. O processo continua até que cada objeto forme um único grupo.

Assim, uma característica importante desses métodos é a propriedade de hierarquia. Visto que esses não exigem a definição de um número de grupos *a priori*, é possível inferir esse número através da análise do gráfico chamado Dendrograma, que apresenta um resumo do histórico do agrupamento.

Alguns algoritmos hierárquicos aglomerativos populares usados para construir agrupamentos são: (1) ligação simples (*single linkage*); (2) ligação completa (*complete linkage*); (3) ligação média (*average linkage*); (4) método de Ward; e (5) método do centróide, dentre outros. Esses algoritmos diferem na maneira através da qual a distância entre os grupos é computada (Corrar et al., p. 346, 2014).

Segundo Hair et al. (2005) a técnica de agrupamento geralmente envolve três passos:

> "O primeiro é a medida de alguma forma de similaridade ou associação entre as entidades para determinar quantos grupos realmente existem na amostra. O segundo é o próprio processo de agrupamento, nas quais as entidades são particionadas em grupos (agrupamento). O último passo é estabelecer o perfil das pessoas ou variáveis para determinar sua composição." (Hair et al., p. 33, 2005)

2.3.1 Método Ward

Em 1963 Ward propôs um método de agrupamento com o intuito de formar grupos com o máximo de homogeneidade interna.

Em cada passo do agrupamento, a soma dos quadrados dos desvios dos valores observados de cada objeto em cada variável, em relação à respectiva média do grupo ao qual o objeto pertence, é minimizada com o objetivo de que seja mínima a variação dentro dos grupos formados (Corrar et al.,2014). Essa soma é chamada de Soma de Quadrados dentro dos grupos formados, ou Soma de Quadrados Residual.

Hair et al. (pag. 34, 2005) explicam que "em cada estágio do procedimento de agrupamento, a soma interna de quadrados é minimizada sobre todas as partições possíveis, que podem ser obtidas pela combinação de dois agregados do estágio anterior."

No primeiro passo do algoritmo cada objeto (elemento amostral) é considerado um grupo e para cada passo subsequente é calculada a soma dos quadrados dentro dos grupos possíveis de serem formados de acordo com o resultado do passo anterior do algoritmo.

De acordo com Mingoti (2005) considerando g* o número de *clusters* em uma partição em determinado momento no algoritmo, sejam:

$X_{ij} = (X_{i1j}X_{i2j} \dots X_{ipj})'$, o vetor de medidas observadas para o j-ésimo elemento amostral do i-ésimo grupo;

$\bar{X}_{i.} = (\bar{X}_{i1.}\bar{X}_{i2.} \dots \bar{X}_{ip.})'$, o vetor de médias (centróide) do i-ésimo grupo C_i;

$\bar{X}$ $(\bar{X}_{.1}\bar{X}_{.2} \dots \bar{X}_{.p})'$, o vetor de médias global, sem levar em conta qualquer partição dos dados amostrais, onde $\bar{X}_{.l} = \frac{1}{n}\sum_{i=1}^{g^*}\sum_{j=1}^{n_i} X_{ilj}$, l = 1,2, ... ,p, n_i o número de elementos no conglomerado C_i quando se está no passo k do processo de agrupamento, *n* o número total de observações amostrais e p o número de variáveis.

Define-se a Soma de Quadrados Total (SST_c) como:

$$SST_c = \sum_{i=1}^{g^*}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})'(X_{ij} - \bar{X}) \qquad (1)$$

A Soma de Quadrados Total dentro dos grupos da partição (Soma de Quadrados Residual) é definida como:

$$SSR = \sum_{i=1}^{g^*} SS_i = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)' (X_{ij} - \bar{X}_{i.}) \quad (2)$$

sendo:

$$SS_i = \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i.})' (X_{ij} - \overline{X}_{i.}) \quad (3)$$

a soma de quadrados residual do i-ésimo conglomerado, i=1,2,...,g*. A soma de Quadrados entre os g* grupos (partição) formados no passo k do algoritmo de agrupamento é dada como:

$$SSB = \sum_{i=1}^{g^*} n_i (\bar{X}_{i.} - \bar{X})' (\bar{X}_{i.} - \bar{X}) \quad (4)$$

A Soma de Quadrados Total é igual a soma: SSB + SSR.

Assim, no método de Ward em cada passo do algoritmo busca-se os dois grupos (dentre os disponíveis para agrupamento), cuja união minimiza a soma de quadrados dentro dos grupos (2). Em 1963 Ward mostrou em seu artigo que esse procedimento é equivalente a unir-se os dois grupos que minimizam a distância definida como em (5), sendo C_l e C_i os grupos que estão sendo comparados.

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i}\right] (\bar{X}_l - \bar{X}_{i.})' (\bar{X}_l - \bar{X}_{i.}) \quad (5)$$

No método de Ward são unidos em cada passo os dois *clusters* com menor valor de (5).

De acordo com Mingoti (2005) este método é similar ao método centróide, porém ele considera os tamanhos dos conglomerados que estão disponíveis para agrupamento em cada passo do algoritmo, dando pesos diferentes para cada possível união.

Segundo Hair et al. (2005) em muitas situações o número de grupos final da partição desejada não é pré-fixado mas deve ser estimado usando os dados amostrais. Para auxiliar na decisão do número de grupos a serem formados, tem-se algumas medidas que podem ser utilizadas como:

- Análise do comportamento do nível de fusão (distância);
- Análise do comportamento do nível de similaridade;
- Análise do coeficiente de correlação intra classe R^2;
- Estatística *Pseudo F*;
- Correlação semiparcial (método de Ward);
- Estatística *Pseudo* T^2, dentre outras.

Nesta monografia serão utilizados o nível de similaridade, o coeficiente de correlação intra-classe e a estatística *Pseudo F* para estimação do número de grupos da partição das empresas. Essas medidas serão descritas na seção a seguir.

2.3.2 *Estimação do Número de Clusters da Partição*

Segundo Mingoti (2005) na análise do comportamento do nível de similaridade verifica-se os pontos nos quais há um decaimento acentuado demonstrando que a similaridade entre os grupos formados diminuiu substancialmente. Neste momento o algoritmo deveria ser interrompido e o número g* relacionado ao passo anterior Lé uma estimativa do número de grupos g da partição final dos dados. Nesta monografia utilizamos o coeficiente de similaridade implementado no *software* Minitab for Windows versão 16 que é definido como:

$$S_{il} = \left(1 - \frac{d_{il}}{\max\{d_{jk}\,, j,k=1,2,\ldots,n\}}\right) x\ 100 \qquad (6)$$

onde $\max\{d_{jk}\,, j,k = 1,2,\ldots,n\}$ é a maior distância entre os *n* elementos amostrais na matriz de distâncias D_{nxn} do primeiro estágio do processo de agrupamento, d_{il} é a distância entre os clusters C_i , C_l , e S_{il} é a similaridade entre esses clusters. É importante observar que o coeficiente de similaridade como definido em (6) pode assumir valores negativos já que a distância entre os conglomerados C_i , C_l

comparados, pode eventualmente ser maior que a distância máxima entre os *n* elementos amostrais.

Para avaliar a qualidade da partição formada com g* grupos utiliza-se o coeficiente de correlação intra-classe (R^2) e a estatística *Pseudo F*. O coeficiente de Correlação Intra-classe (R^2), é definido como:

$$R^2 = \frac{SSB}{SST_c} \tag{7}$$

e representa a proporção da variabilidade total dos dados que é explicada pela partição em g* grupos realizada nos dados. Quanto maior for o valor do R^2 melhor é a partição, pois maior será a soma de quadrados entre grupos SSB e menor será o valor da soma de quadrados residual SSR.

A estatística *Pseudo F* é definida em (8). Quanto maior for o seu valor melhor é a partição formada, pois indica que os *clusters* possuem vetores de médias bem distintos.

$$PF = \left(\frac{n-g^*}{g^*-1}\right)\left(\frac{R^2}{1-R^2}\right) \tag{8}$$

A estatística *Pseudo F* é uma função do coeficiente R^2, mas leva em consideração o número de grupos o que não é feito no coeficiente R^2.

Segundo Mingoti (2005) é importante destacar que o maior valor possível de R^2 é 1 que ocorre na partição que tem *n* conglomerados, ou seja cada elemento da amostra é um conglomerado. Desta forma, o valor de R^2 precisa ser analisado com cuidado. Em geral, utiliza-se o coeficiente de similaridade para definir uma região de possíveis valores (estimativas) para o número de *clusters* g. Nessa região o valor de R^2 é analisado juntamente com o valor de *Pseudo F*.

2.3.3 Método das K-Médias

No método das K-médias "cada elemento amostral é alocado àquele *cluster* cujo centróide (vetor de médias amostral) é o mais próximo do vetor de valores observados para o respectivo elemento. " (Mingoti, pag. 192, 2005).

Segundo Corrar et al. (2014) o método das K-Médias é um procedimento também conhecido como método de partição, pois busca uma partição de *n* objetos preservando a homogeneidade dentro dos grupos e a heterogeneidade entre os grupos.

O método das K-médias se resume em dois passos, o primeiro é especificar as sementes iniciais do agrupamento; o segundo é alocar cada uma das observações amostrais a uma das sementes definidas, com base na similaridade. Existem inúmeros métodos para auxiliar na designação das sementes, a ideia é que cada observação amostral se agrupe no grupo cuja semente seja mais parecida com ela. (Hair et al., 2005)

Johnson e Wichern (2007) afirmam que há fortes razões para que as sementes não sejam previamente fixadas, pois se as sementes informadas forem semelhantes haverá pouca diferenciação nos grupos formados.

De acordo com Mingoti (2005) as escolhas das sementes iniciais influenciam fortemente na formação do agrupamento final. Desta forma, tem-se algumas sugestões para auxiliar nesta decisão: (i) utilização de uma técnica hierárquica aglomerativa para construir os grupos iniciais, sendo que o vetor de médias de cada grupo representará uma semente inicial para o método das K-Médias; (ii) escolher as sementes através da escolha da variável aleatória de maior variância; divide-se o domínio da variável em k intervalos e a semente inicial será o centróide de cada intervalo (MESQUITA, 2010); (iii) buscar os k elementos mais discrepantes no conjunto de dados, cada elemento podendo ser uma semente inicial, etc.. A escolha de sementes através de um processo aleatório a partir de uma amostragem aleatória simples sem reposição de k elementos amostrais do banco de dados, não é considerado um método eficiente.

De acordo com Hair et al. (2005) apesar de não recomendado, o pesquisador pode escolher as sementes a serem utilizadas, se o objetivo for a validação de uma solução já existente.

Grande parte dos *softwares* estatísticos utilizam como *default* as k primeiras observações no banco de dados como sementes iniciais de agrupamento. Caso esses primeiros elementos sejam similares entre si, o uso desse procedimento não é recomendado, sendo mais adequado que o pesquisador especifique quais sementes deseja utilizar para inicialização do algoritmo (Mingoti, 2005)

3 Descrição das Variáveis e Banco de Dados

Para a pesquisa tratada nessa monografia foram utilizados dados de uma empresa real, portanto tem-se um estudo de caso. Define-se como estudo de caso a pesquisa "sobre um determinado indivíduo, família, grupo ou comunidade que seja representativo do seu universo, para examinar aspectos variados de sua vida." (CERVO; BERVIAN, 2004, p. 67).

Foi utilizada a técnica estatística de análise de agrupamento, primeiro através do método de Ward com o uso da análise de similaridade, do coeficiente de correlação intra-classe (R^2), bem como a estatística *Pseudo F* para definir as possíveis estimativas para o número g de grupos e a qualidade estatística da partição relacionada.

Posteriormente, utilizou-se o método das K-Médias com o número de grupos k=g previamente estabelecido, para refinar a solução obtida no método de Ward ou validá-la.

Para elaboração do trabalho foi utilizada a carteira de clientes de uma consultoria atuarial composta de trinta e três operadoras de planos de saúde. São operadoras de várias regiões do país, empresas de pequeno, médio e grande porte. Neste estudo 55% das operadoras são classificadas como medicina de grupo, 24% como autogestões, 12% filantropia e 9% como cooperativas médicas.

A escolha dessas trinta e três operadoras foi devido ao fato de elas contemplarem a carteira de uma consultoria atuarial do mercado brasileiro. Este estudo será disponibilizado para que esta empresa possa utilizar as informações aqui obtidas para melhorar a gestão de sua carteira de clientes. Os nomes e números de registros das operadoras não foram informados para que o sigilo das mesmas fosse preservado.

Para alocar as operadoras pela técnica de conglomerados foram consideradas as variáveis despesa assistencial, despesa administrativa, receita de contraprestação e beneficiários, definidas conforme o Quadro 1. Estas variáveis foram selecionadas porque são informações relevantes no mercado de saúde e de envio obrigatório à Agência Nacional de Saúde Suplementar (ANS).

Quadro 1 - Descrição das variáveis utilizadas no estudo

VARIÁVEL	CONCEITO
Despesa Assistencial	A soma do gasto de toda e qualquer utilização, pelo beneficiário, das coberturas contratadas, referente a prestação direta dos serviços de assistência à saúde.
Despesa Administrativa	A soma dos gastos que não são referentes a prestação direta dos serviços de assistência à saúde.
Receita de Contraprestações	A soma dos valores que é pago à operadora pelo beneficiário para a prestação dos serviços de assistência à saúde.
Beneficiários	A variável beneficiários refere-se a quantidade de pessoas que possui contrato assinado com a operadora de plano de saúde para garantia de assistência médico-hospitalar ou odontológica.

Fonte: Base de dados em estudo

Os valores apurados são anuais referentes ao período de janeiro a dezembro de 2013, para as variáveis despesa assistencial, despesa administrativa e receita de contraprestação. Já para a variável beneficiários considerou-se a média mensal no ano de 2013, para esta variável não se considerou a soma de beneficiários no ano, para que o valor obtido não ficasse superestimado, uma vez que os beneficiários tendem a permanecer vinculado ao plano de saúde por meses ou até anos.

Os dados utilizados nessa monografia são de acesso público para consulta, disponibilizados no site da Agência Nacional de Saúde Suplementar e se encontram no Anexo I dessa monografia.

4 Análise Estatística dos Dados

Nesta seção apresenta-se uma análise descritiva dos dados e os resultados obtidos com a aplicação do método de agrupamento de Ward, buscando analisar e demonstrar as principais vantagens oferecidas pela sua aplicação.

No Quadro 2 encontram-se as algumas estatísticas descritivas das variáveis estudadas.

Quadro 2 - Estatísticas descritivas das variáveis (*n*=33 operadoras)

Variáveis	Receita de Contraprestação	Despesa Administrativa	Despesa Assistencial	Beneficiários
Soma (Total)	1.012.482.571,00	256.028.057,00	813.760.165,00	617.473,00
Média	30.681.290,03	7.758.425,97	24.659.398,94	18.711,30
Desvio Padrão	41.125.639,49	10.091.299,01	34.723.674,20	21.003,94
Mínimo	576.115,00	265.251,00	195.336,00	209,00
Máximo	167.671.337,00	40.502.169,00	149.115.380,00	72.649,00
Mediana	11.262.962,00	3.537.879,00	9.026.922,00	10.432,00
Coeficiente de Variação	134,04	130,07	140,81	112,25
Assimetria	1,91	1,87	2,07	1,38
Curtose	3,64	3,02	4,65	1,01

Fonte: Base de dados em estudo

As trinta e três operadoras analisadas obtiveram uma receita total de R$1.012.482.571,00 no ano de 2013, uma média de R$ 30.681.290,03 por operadora. A receita mínima encontrada foi de R$ 576.115,00 e a máxima de 167.671.337,00, o desvio-padrão apurado foi de R$ 41.125.639,49 e a mediana de R$ 11.262.962,00.

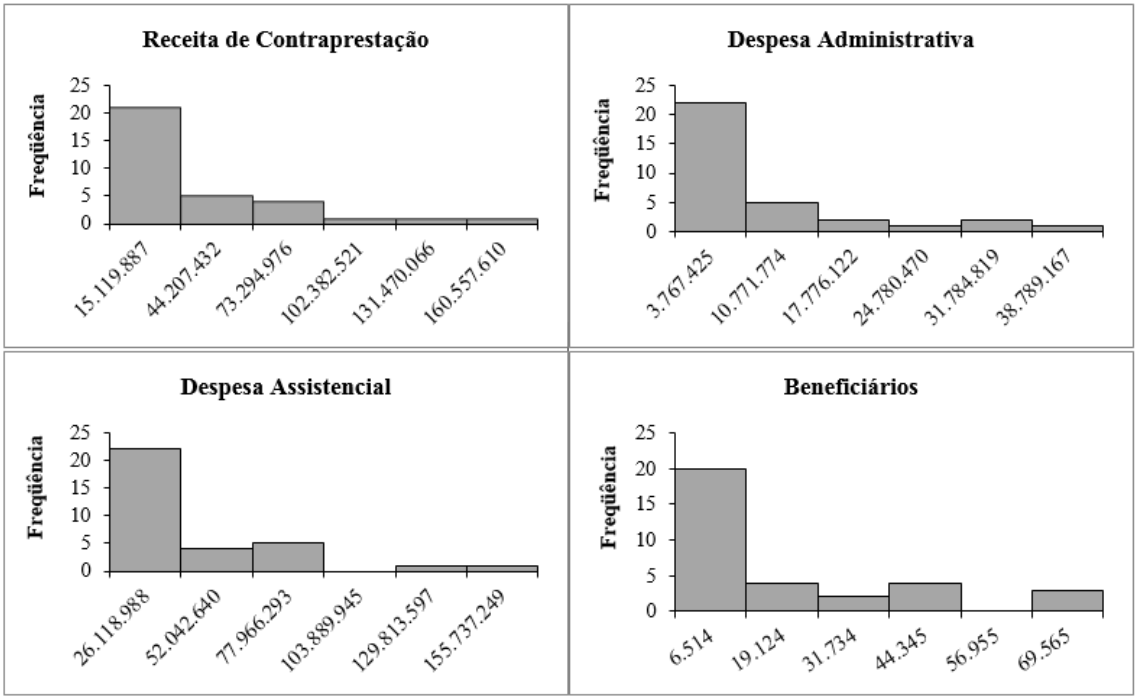
A despesa administrativa total foi de R$ 256.028.057,00 no ano de 2013, uma média de R$ 7.758.425,97 por operadora. A despesa administrativa mínima encontrada foi de R$ 265.251,00 e a máxima de R$ 40.502.169,00, o desvio-padrão apurado foi de R$ 10.091.299,01 e a mediana de R$ 3.537.879,00.

A despesa assistencial total foi de R$ 813.760.165,00 no ano de 2013, uma média de R$ 24.659.398,94 por operadora. A despesa assistencial mínima foi de 195.336,00 e a máxima de R$ 149.115.380,00, o desvio-padrão apurado foi de R$ 34.723.674,20 e a mediana de R$ 9.026.922,00.

Tem-se um total de 617.473 beneficiários no ano de 2013, uma média de 18.711,30 por operadora. A operadora com menor quantidade de beneficiários possui 209 pessoas e a operadora com a maior quantidade 72.649, um desvio-padrão de 21.003,94 e uma mediana de 10.432.

Para comparar a dispersão das variáveis em estudo utilizou-se o coeficiente de variação. A variável com maior coeficiente de variação é a despesa assistencial e a menor é beneficiários. Todas as variáveis em estudo possuem uma distribuição assimétrica positiva e distribuição *leptocúrtica*, ou seja, a medida de curtose é maior que o da distribuição normal.

Analisando os histogramas apresentados na Figura 1, é possível observar melhor a distribuição assimétrica para todas as variáveis analisadas.

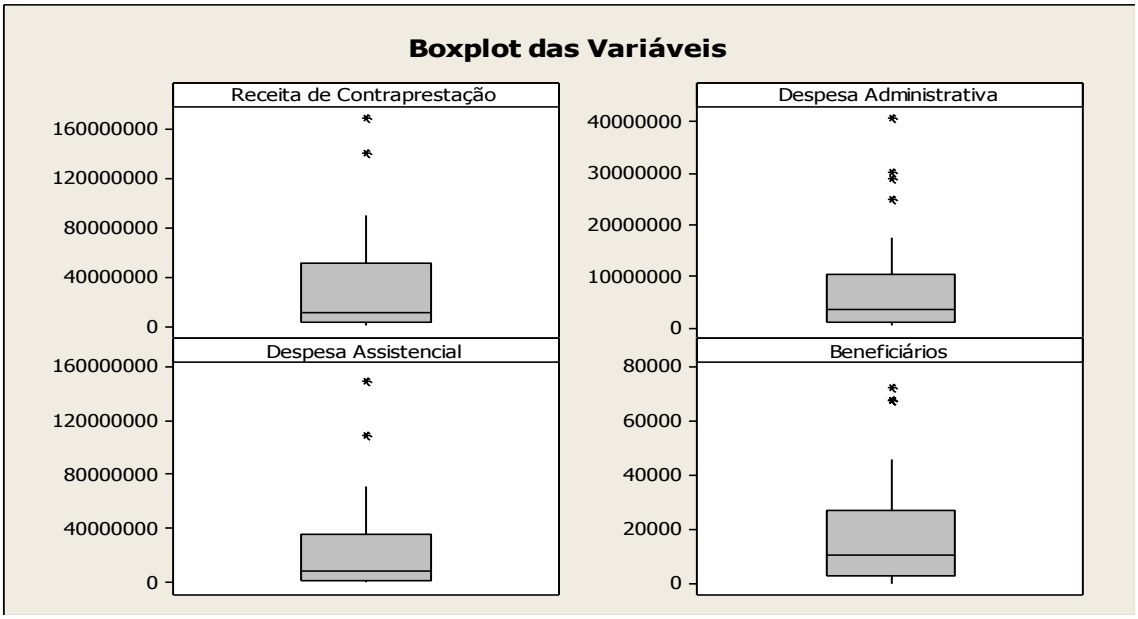


Fonte: Base de dados em estudo

Figura 1 - Histograma das Variáveis

Nos gráficos Box-Plots da Figura 2, é possível observar a presença de alguns valores discrepantes. A operadora 30 que aparece como *outlier* no boxplot de todas as variáveis, é uma operadora de grande porte no mercado de saúde brasileiro, por isso apresenta valores tão elevados. A operadora 24 apresenta *outliers* nas variáveis receita e despesa assistencial, também é uma operadora de grande porte. Um dos valores discrepantes da variável beneficiários é da operadora 17; ela possui muitos beneficiários por ser a principal operadora de plano de saúde no estado em que atua. Na despesa administrativa os *outliers* foram os dados das operadoras 1, 5, 4 e 30. Com exceção da operadora 30 as demais operadoras são classificadas como autogestão, é comum que esse tipo de operadora tenha valores elevados de despesa administrativa em comparação com o mercado.

Diante do exposto foi possível verificar que os valores discrepantes encontrados nos dados não apresentam inconsistência das informações, são na verdade características reais das operadoras em estudo.



Fonte: Base de dados em estudo

Figura 2 - Boxplot das Variáveis

Segundo Corrar et al. (2014) a maioria das medidas de similaridade são sensíveis a diferentes escalas ou magnitudes entre os dados. Visto que nem todas as variáveis em estudo estão na mesma escala e que há uma grande variação entre os dados se optou por utilizar a padronização com o intuito de evitar inconsistências nas soluções apuradas. A padronização dos dados foi realizada através da divisão do valor observado de cada variável pelo desvio-padrão amostral correspondente. Os valores padronizados podem ser encontrados no Anexo I dessa monografia.

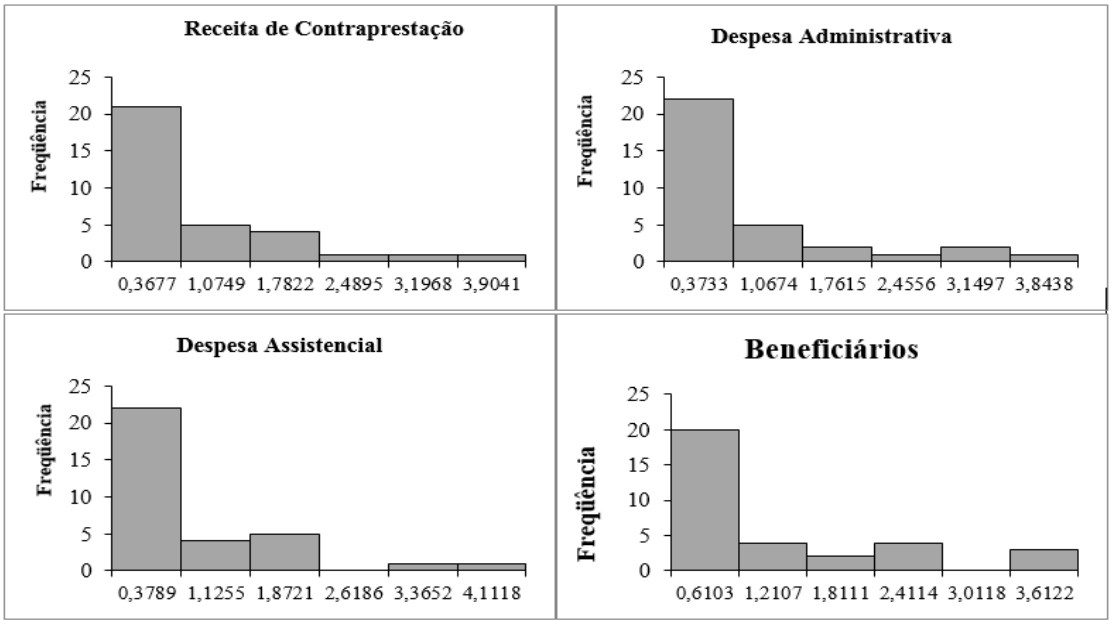
No Quadro 3 pode ser observado as estatísticas descritivas das variáveis após a padronização.

Quadro 3 - Estatísticas descritivas das variáveis padronizadas (*n*=33 operadoras)

Variável	Receita de Contraprestação	Despesa Administrativa	Despesa Assistencial	Beneficiários
Soma (Total)	24,62	25,37	23,44	29,40
Média	0,75	0,77	0,71	0,89
Desvio Padrão	1,00	1,00	1,00	1,00
Mínimo	0,01	0,03	0,01	0,01
Máximo	4,08	4,01	4,29	3,46
Mediana	0,27	0,35	0,26	0,50

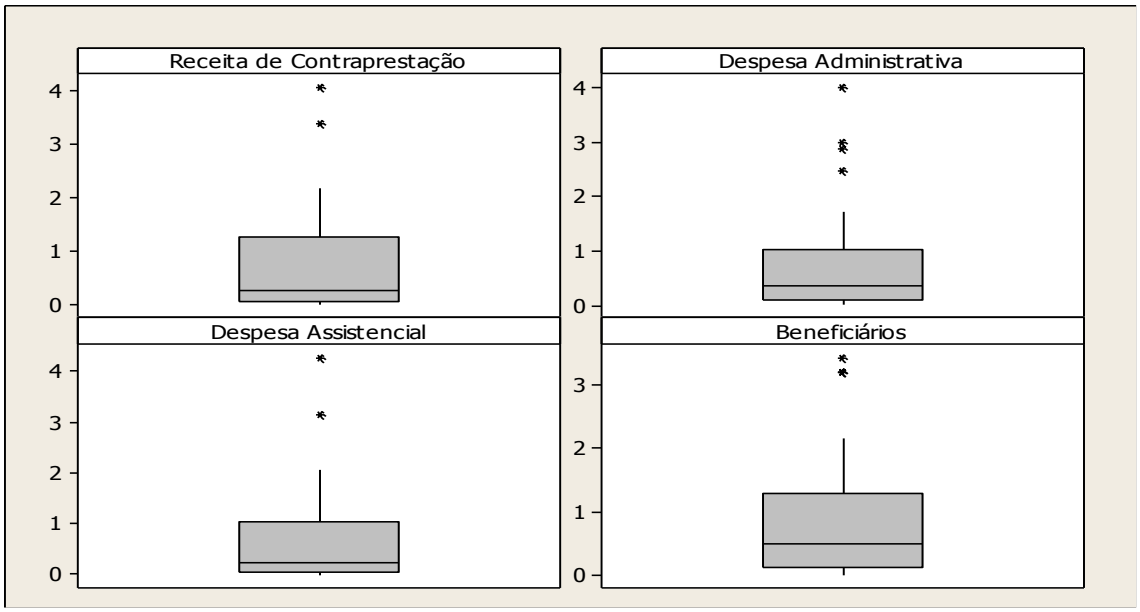
Fonte: Base de dados em estudo

Após a padronização torna-se possível a comparação entre as variáveis. Verifica-se que a variável beneficiários possui a maior soma e média, seguida pela variável despesa administrativa, receita de contraprestação e por último a despesa assistencial. O valor mínimo é igual para as variáveis receita de contraprestação, despesa assistencial e beneficiários (0,01), já para a despesa administrativa é de 0,03. Devido a padronização o desvio-padrão de todas as variáveis em estudo é igual a 1,00. Na Figura 3 observa-se a distribuição assimétrica das variáveis padronizadas Na Figura 4, apresenta-se o boxplot das variáveis padronizadas. Devido a padronização os valores das variáveis se alteraram, porém, o formato do gráfico permaneceu semelhante aos gráficos das variáveis originais. Os pontos de *outliers* correspondem as mesmas operadoras mencionadas na Figura 2.



Fonte: Base de dados em estudo

Figura 3 - Histograma das Variáveis Padronizadas

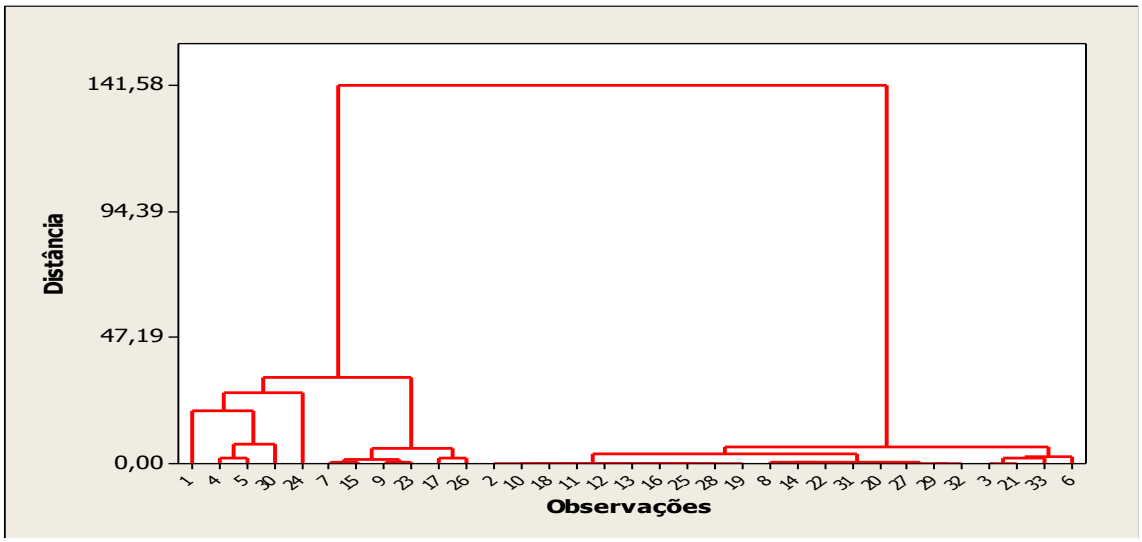


Fonte: Base de dados em estudo

Figura 4 - Boxplot das Variáveis Padronizadas

4.1 Análise de agrupamento pelo método Ward

Como dito anteriormente os procedimentos hierárquicos aglomerativos formam uma construção similar a uma árvore denominada dendrograma. Através da utilização do método de Ward considerando a distância definida em (5) na seção 2.3.1, obtém-se o dendrograma gerado pelo *software* utilizado Minitab for Windows 16 para os dados padronizados, apresentado na Figura 5. Os resultados do histórico, passo-a-passo, do agrupamento se encontram no Quadro 4.



Fonte: Base de dados em estudo

Figura 5 - Dendograma: agrupamento das operadoras

Quadro 4 - Histórico do agrupamento das 33 operadoras - Método Ward

Passo	Número de clusters	Nível de similaridade	Nível de Distância	Grupos Unidos
1	32	99,999	0,001	16 e 25
2	31	99,990	0,004	10 e 18
3	30	99,989	0,005	12 e 13
4	29	99,978	0,009	Grupos unidos no passo nº 2 e 11
5	28	99,965	0,014	Grupos unidos no passo nº 1 e 28
6	27	99,965	0,014	29 e 32
7	26	99,943	0,023	Grupos unidos nos passos nº 4 e nº 3
8	25	99,941	0,024	20 e 27
9	24	99,885	0,047	Grupos unidos nos passos nº 5 e nº 7
10	23	99,883	0,048	22 e 31
11	22	99,832	0,069	Grupos unidos no passo nº 9 e 19
12	21	99,800	0,082	Grupos unidos no passo nº 10 e 14
13	20	99,658	0,141	3 e 21
14	19	99,509	0,202	Grupos unidos nos passos nº 6 e nº 8
15	18	99,427	0,236	Grupos unidos no passo nº 11 e 2
16	17	98,908	0,449	Grupos unidos no passo nº 12 e 8
17	16	98,884	0,459	9 e 23
18	15	98,277	0,708	Grupos unidos nos passos nº 14 e nº16
19	14	98,223	0,73	7 e 15
20	13	95,946	1,666	Grupos unidos nos passos nº 17 e nº 19
21	12	95,405	1,889	Grupos unidos no passo nº 13 e 33
22	11	95,108	2,011	17 e 26
23	10	94,878	2,105	4 e 5
24	9	94,292	2,346	Grupos unidos no passo nº 21 e 6
25	8	91,711	3,407	Grupos unidos nos passos nº 15 e 18
26	7	85,911	5,791	Grupos unidos nos passos nº 20 e nº 22
27	6	84,175	6,504	Grupos unidos nos passos nº 24 e nº 25
28	5	82,778	7,079	Grupos unidos no passo nº 23 e 30
29	4	52,365	19,579	Grupos unidos no passo nº 28 e 1
30	3	35,081	26,684	Grupos unidos no passo nº 29 e 24
31	2	21,922	32,092	Grupos unidos nos passos nº 26 e nº 30
32	1	-244,457	141,581	Grupos unidos nos passos nº 27 e nº 31

Fonte: Base de dados em estudo

Através da análise do nível de similaridade observou-se um decaimento mais acentuado do passo 28 para o passo 29. A similaridade de 82,778% cai para 52,365% e a distância salta de 7,079 para 19,579, o que indica que o algoritmo deve ser interrompido no passo 28. Os valores dos níveis de similaridade são semelhantes nos passos 26 a 28 (entre 86% e 83%), assim a região k = 5, 6, 7 deveria ser pesquisada. Por parcimônia o valor mais indicado na região seria aquele relacionado com o passo 28, ou seja, k=5. Apesar de terem níveis de similaridade baixos, para efeito desse estudo decidiu-se explorar também as soluções com k=3 e k=4.

No Quadro 5 encontram-se os valores do coeficiente de correlação intra-classe (R^2) e a estatística Pseudo *F* das partições na região de número de grupos (k) analisada.

Quadro 5 - Soma de quadrados das Partições, R^2 e Pseudo *F*

K (no. de grupos)	SSR	SSB	R^2 %	*Pseudo F*
3	41,16	86,84	67,84	31,64
4	27,82	100,18	78,26	34,81
5	18,03	109,97	85,91	42,69
6	14,49	113,51	88,68	42,29
7	11,24	116,76	91,22	45,01

Soma de Quadrados Total (SSTc) = 128

Fonte: Base de dados em estudo

Observa-se que o SSR diminui e o R^2 fica maior à medida que o número de grupos aumenta, como esperado. Como melhores resultados se apresentam k=6 e k=7, com as menores variabilidades dentro do agrupamento e maiores percentual de explicação dos grupos formados, mesmo sendo o resultado do *Pseudo F* para k=6 inferior ao resultado obtido em k=5.

Nos Quadros 6-10 apresentam-se os agrupamentos formados pelo método Ward para cada um dos valores de número de grupos k do Quadro 5, com os respectivos valores de Soma de Quadrados dentro dos grupos.

A partição com k=3 (Quadro 6) indica um primeiro grupo composto por cinco operadoras, o segundo composto por vinte e duas e o terceiro por seis, sendo que o primeiro tem um valor de Soma de Quadrados Residual bem maior, embora não seja o composto pelo maior número de elementos. Este grupo com a maior soma de quadrados residual é constituído pelas operadoras apontadas como *outliers*.

Quando um quarto grupo é adicionado (ver Quadro 7), uma única operadora passa a compor este grupo, 24, que anteriormente fazia parte do grupo número um. A retirada desta operadora do grupo de pertinência da solução k=3, faz com que a respectiva SSR do grupo 1 diminua seu valor pela metade.

Na formação com k=5 (ver Quadro 8), a operadora 1 passa a compor um novo grupo e pode-se visualizar uma redução considerável no SSR do grupo ao qual ela pertencia (solução com k=4). Já na formação de 6 grupos (Quadro 9) a operadora 30 passa a compor um novo aglomerado e pode-se visualizar uma redução considerável no SSR do grupo ao qual ela pertencia. Na inclusão de um sétimo grupo (Quadro 10) as operadoras 3, 6, 21 e 33 que compunham em todos os agrupamentos anteriores o grupo de número 2 passam a formar um novo aglomerado.

Observou-se que o método Ward inicialmente separa em um grupo as operadoras apontadas como *outliers* na análise descritiva (1, 4, 5 e 30) e a medida que a quantidade de grupos aumenta essas operadoras vão se separando, exceto pelas empresas 4 e 5 que permanecem juntas em todos os agrupamentos formados neste estudo.

Existem dois grupos bem definidos um com seis operadoras (7, 9, 15, 17, 23 e 26) e o outro com vinte e duas (2, 3, 6, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33). O grupo com seis operadora permanece presente em todos os agrupamentos e o segundo se divide somente na análise de k=7 onde algumas das operadoras se separam para formar um novo grupo (3, 6, 21 e 33).

Quadro 6 - Agrupamento Ward - k=3

Grupos	Operadoras	SSR
n1 = 5	1, 4, 5, 24 e 30	**27,7234**
n2 = 22	2, 3, 6, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**8,1114**
n3 = 6	7, 9, 15, 17, 23 e 26	**5,3287**

Fonte: Base de dados em estudo

Quadro 7 - Agrupamento Ward - k=4

Grupos	Operadoras	SSR
n1 = 4	1, 4, 5 e 30	**14,3816**
n2 = 22	2, 3, 6, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**8,1114**
n3 = 6	7, 9, 15, 17, 23 e 26	**5,3287**
n4 = 1	24	**0,0000**

Fonte: Base de dados em estudo

Quadro 8 - Agrupamento Ward - k=5

Grupos	Operadoras	SSR
n1 = 1	1	**0,0000**
n2 = 22	2, 3, 6, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**8,1114**
n3 = 3	4, 5 e 30	**4,5920**
n4 = 6	7, 9, 15, 17, 23 e 26	**5,3287**
n5 = 1	24	**0,0000**

Fonte: Base de dados em estudo

Quadro 9 - Agrupamento Ward - k=6

Grupos	Operadoras	SSR
n1 = 1	1	**0,0000**
n2 = 22	2, 3, 6, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**8,1114**
n3 = 2	4 e 5	**1,0526**
n4 = 6	7, 9, 15, 17, 23 e 26	**5,3287**
n5 = 1	24	**0,0000**
n6 = 1	30	**0,0000**

Fonte: Base de dados em estudo

Quadro 10 - Agrupamento Ward - k=7

Grupos	Operadoras	SSR
n1 = 1	1	**0,0000**
n2 = 18	2, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 22, 25, 27, 28, 29, 31 e 32	**2,6716**
n3 = 4	3, 6, 21 e 33	**2,1876**
n4 = 2	4 e 5	**1,0526**
n5 = 6	7, 9, 15, 17,23 e 26	**5,3287**
n6 = 1	24	**0,0000**
n7 = 1	30	**0,0000**

Fonte: Base de dados em estudo

4.2 Análise de agrupamento pelo método das K-médias

A seguir apresenta-se os resultados obtidos do método K-Médias quando se utiliza como sementes iniciais os centróides provenientes dos grupos formados pelo método de Ward. A composição dos grupos para cada valor de k é mostrada nos Quadros 11-15.

Para três grupos os conglomerados obtidos foram muito próximos aos resultantes do método de Ward, observa-se que igualmente a este método nesta análise as operadoras com valores discrepantes também foram alocadas em um único grupo. A

única diferença foi a operadora 6 que no método Ward pertencia ao grupo dois e agora foi alocada no grupo três.

Já para os resultados obtidos para quatro grupos houve alteração em todos os conglomerados. A operadora 1 passa a formar um grupo, o grupo dois perde a operadora 6 que passa a ser alocada no grupo 3. Esse por sua vez, perde a operadora 26 para o grupo quatro e este grupo passa a ser composto também pelas operadoras que antes estavam no primeiro grupo (4, 5 e 30).

Em cinco grupos a operadora 6 saiu do grupo 2 para compor um grupo com algumas das operadoras que formavam o agrupamento quatro no método Ward. A operadora 24 que anteriormente compunha um conglomerado passa a fazer parte de um grupo com mais três outras operadoras. As operadoras 5 e 30 que formavam um agrupamento juntamente com a operadora 4 passam a compor um conglomerado. Destaca-se que para cinco grupos os métodos Ward e das K-médias apresentam significativas mudanças na formação de seus agrupamentos.

Para seis grupos os conglomerados nos dois métodos estudados são bem próximos. Cita-se como diferença a operadora 6 que no método Ward compunha o grupo 2 e no método das K-médias está localizada no grupo 4.

Para sete aglomerados os dois métodos também formam grupos bem similares, neste caso cita-se como diferença a operadora 8 que está alocado no grupo três e no método Ward fazia parte do grupo dois.

O método das K-médias possui um grupo bem definido composto de vinte e uma operadoras (2, 3, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33), elas permanecem como um conjunto em quase todos os agrupamentos só se desmembrado quando são formadas sete partições. Sendo muito similar ao grupo de vinte e duas operadoras no método Ward (2, 3, 6, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33) que também só apresenta modificações no último agrupamento (k=7).

Quadro 11- Agrupamento K-Médias - k=3

Grupos	Operadoras	SSR
n1 = 5	1, 4, 5, 24 e 30	**27,7234**
n2 = 21	2, 3, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**5,2170**
n3 = 7	6, 7, 9, 15, 17, 23 e 26	**7,3090**

Fonte: Base de dados em estudo

Quadro 12 - Agrupamento K-Médias - k=4

Grupos	Operadoras	SSR
n1 = 1	1	**0,0000**
n2 = 21	2, 3, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**5,2170**
n3 =5	6, 7, 9, 15 e 23	**2,6520**
n4 =6	4, 5, 17, 24, 26 e 30	**22,5560**

Fonte: Base de dados em estudo

Quadro 13- Agrupamento K-Médias - k=5

Grupos	Operadoras	SSR
n1 = 1	1	**0,0000**
n2 = 21	2, 3, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**5,2170**
n3 = 5	6, 7, 9, 15 e 23	**2,6520**
n4 = 4	4, 17, 24 e 26	**14,9290**
n5 = 2	5 e 30	**2,1660**

Fonte: Base de dados em estudo

Quadro 14 - Agrupamento K-Médias - k=6

Grupos	Operadoras	SSR
n1 = 1	1	**0,0000**
n2 = 21	2, 3, 8, 10, 11, 12, 13, 14, 16, 18, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32 e 33	**5,2170**
n3 = 2	4 e 5	**1,0530**
n4 = 7	6, 7, 9, 15, 17, 23 e 26	**7,3090**
n5 = 1	24	**0,0000**
n6 = 1	30	**0,0000**

Fonte: Base de dados em estudo

Quadro 15 - Agrupamento K-Médias - k=7

Grupos	Operadoras	SSR
n1 = 1	1	**0,0000**
n2 = 17	2, 10, 11, 12, 13, 14, 16, 18, 19, 20, 22, 25, 27, 28, 29, 31 e 32	**1,9700**
n3 = 5	3, 6, 8, 21 e 33	**2,7160**
n4 = 2	4 e 5	**1,0530**
n5 = 6	7, 9, 15, 17,23 e 26	**5,3290**
n6 = 1	24	**0,0000**
n7 = 1	30	**0,0000**

Fonte: Base de dados em estudo

No Quadro 16 tem-se as estatísticas de soma de quadrados residual (SSR), coeficiente de Correlação intra-classe (R^2) e *Pseudo F* (PF) obtidos para ambos os métodos de análise de *cluster* utilizados.

Quadro 16 - Resultados por método de análise de *cluster*

k	Ward			K-Médias		
	SSR	**R^2 (%)**	**PF**	**SSR**	**R^2 (%)**	**PF**
3	41,16	67,84	31,64	40,25	68,56	32,70
4	27,82	78,26	34,81	30,43	76,23	31,00
5	18,03	85,91	42,69	24,96	80,50	28,89
6	14,49	88,68	42,29	13,58	89,39	45,50
7	11,24	91,22	45,01	11,07	91,35	45,78

Fonte: Base de dados em estudo

Utilizando o método do K-médias tendo como sementes de inicialização o agrupamento formado previamente pelo método de Ward, em nenhum dos sete resultados obtidos a composição dos grupos apareceram totalmente iguais nas duas soluções. Destaca-se que para k=3, k=6 e k=7 os resultados são mais similares; para os dois últimos citados somente as operadoras 6 e 8 não permaneceram no grupo apontado no método de Ward. Já para k=4 e k=5 tem-se formações de agrupamentos bem diferentes, com exceção do grupo que contém a maioria das operadoras (com 22 operadoras no método de Ward e 21 no método K-Médias) que é bem semelhante nas duas soluções.

Analisando os resultados do SSR, R^2 e da estatística Pseudo F, os agrupamentos com k=6 e k=7 apresentam os melhores resultados: baixo valor de SSR, alto percentual de explicação da variância total dos dados pela partições formadas e alto valor de Pseudo F. Desta forma, será analisado a seguir, o perfil dos grupos formados através destes agrupamentos para os dois métodos (Quadros 17-19).

Comparação Ward e K-Médias para k=6 grupos

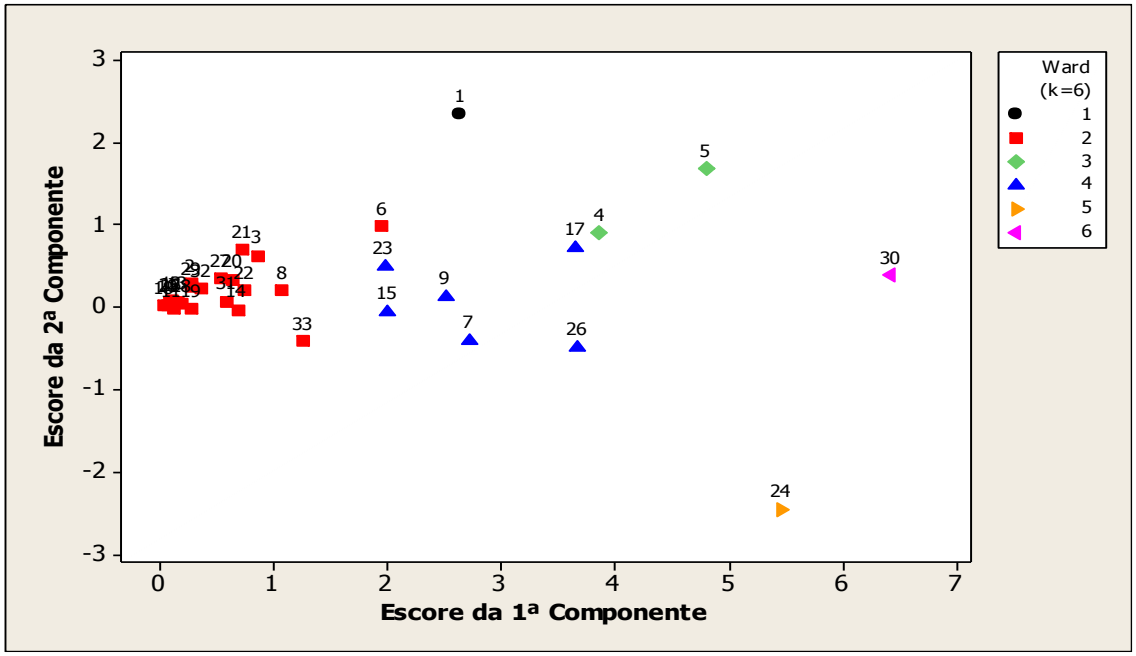
A Análise de Componentes Principais (ACP) é um método da análise multivariada "seu objetivo principal é o de explicar a estrutura de variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, através da construção de combinações lineares das variáveis originais. " (Mingoti, p. 59, 2005)

Segundo Neto e Moita (1997) este método reduz um conjunto de dados que contenham muitas variáveis por um conjunto menor e "embora a informação estatística presente nas n-variáveis originais seja a mesma dos n componentes principais, é comum

obter em apenas 2 ou 3 das primeiras componentes principais mais que 90% desta informação.” (Neto; Moita,1997, p.468)

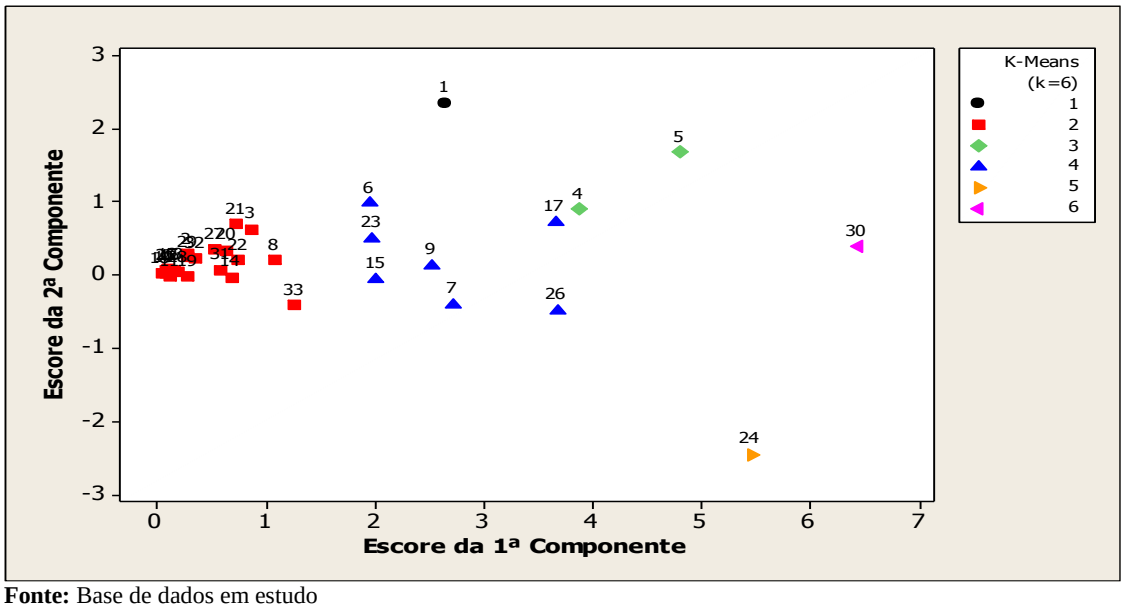
Para auxiliar na análise gráfica dos grupos formados apresentou-se uma análise via componentes principais, com a elaboração de um gráfico de dispersão identificando as soluções obtidas para cada um dos métodos estudados. Destaca-se que foram utilizados os escores das duas componentes principais que explicam 88,9% da variação total (Figuras 6-7, 10-11).

Nas Figura 6 e 7 tem-se os gráficos de dispersão dos dados via ACP identificando cada agrupamento formado. Observa-se que esta análise sugere que a operadora 6 é melhor agrupada no grupo quatro (7, 9, 15, 17, 23 e 26) quando analisamos a proximidade com os elementos do grupo, agrupamento sugerido no método das K-médias.



Fonte: Base de dados em estudo

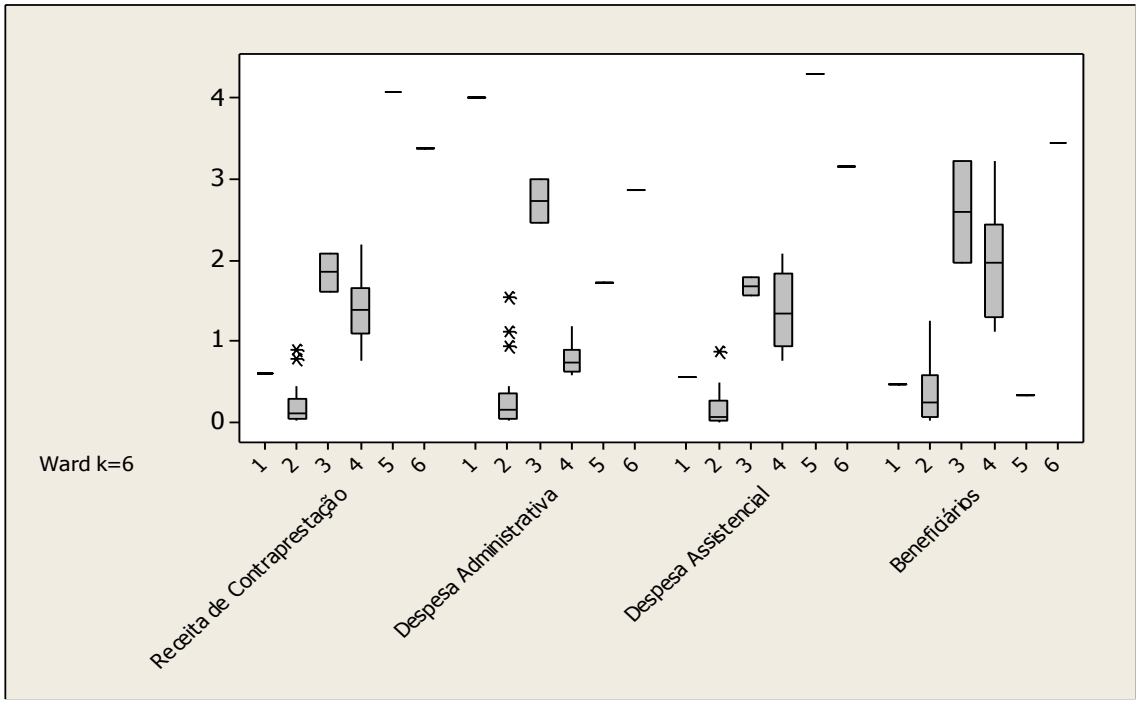
Figura 6 - Escore das duas componentes principais, Ward (k=6)



Fonte: Base de dados em estudo

Figura 7- Escore das duas componentes principais, K-médias (k=6)

Na Figura 8 observa-se o comportamento das variáveis para o agrupamento pelo método Ward para a composição de seis grupos. De um modo geral, observa-se que os grupos formados têm comportamentos diferenciados em todas as variáveis.



Fonte: Base de dados em estudo

Figura 8 - Boxplot das variáveis padronizadas e agrupadas, Ward k=6

Quadro 17 - Estatística Descritiva das variáveis padronizadas, Ward k=6

Ward k=6		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
nº de operadora		1	22	2	6	1	1
Receita de Contraprestação	Média	0,6060	0,2030	1,8505	1,3965	4,0771	3,3911
	Desvio-Padrão	-	0,2402	0,3302	0,4691	-	-
	Q_1	-	0,0372	-	1,0850	-	-
	Mediana	0,6060	0,1143	1,8510	1,3830	4,0771	3,3911
	Q_3	-	0,2885	-	1,6580	-	-
	Mínimo	0,6060	0,0140	1,6170	0,7479	4,0771	3,3911
	Máximo	0,6060	0,8974	2,0840	2,1860	4,0771	3,3911
Despesa Administrativa	Média	4,0136	0,3027	2,7324	0,7742	1,7180	2,8706
	Desvio-Padrão	-	0,3955	0,3753	0,2173	-	-
	Q_1	-	0,0503	-	0,6133	-	-
	Mediana	4,0136	0,1595	2,7320	0,7266	1,7180	2,8706
	Q_3	-	0,3593	-	0,8936	-	-
	Mínimo	4,0136	0,0263	2,4670	0,5876	1,7180	2,8706
	Máximo	4,0136	1,5533	2,9978	1,1869	1,7180	2,8706
Despesa Assistencial	Média	0,5567	0,1729	1,6795	1,3779	4,2943	3,1544
	Desvio-Padrão	-	0,2131	0,1729	0,5033	-	-
	Q_1	-	0,0257	-	0,9290	-	-
	Mediana	0,5567	0,0698	1,6800	1,3480	4,2943	3,1544
	Q_3	-	0,2613	-	1,8340	-	-
	Mínimo	0,5567	0,0056	1,5572	0,7575	4,2943	3,1544
	Máximo	0,5567	0,8712	1,8018	2,0731	4,2943	3,1544
Beneficiários	Média	0,4702	0,3696	2,5978	1,9676	0,3358	3,4588
	Desvio-Padrão	-	0,3560	0,8791	0,7383	-	-
	Q_1	-	0,0707	-	1,2910	-	-
	Mediana	0,4702	0,2325	2,5980	1,9670	0,3358	3,4588
	Q_3	-	0,5784	-	2,4450	-	-
	Mínimo	0,4702	0,0100	1,9762	1,1160	0,3358	3,4588
	Máximo	0,4702	1,2451	3,2193	3,2202	0,3358	3,4588

Fonte: Base de dados em estudo

Analisando a Figura 6 e o Quadro 17 pode-se descrever o agrupamento formado:

Grupo 1: Composto somente pela operadora 1, grupo caracterizado por ter a maior média de gasto com despesa administrativa, possui valores baixos para as demais variáveis.

Grupo 2: Composto por vinte e duas operadoras, possui as menores médias para as variáveis receita de contraprestação, despesa administrativa e despesa assistencial.

Apesar de não apresentar a menor média para a variável beneficiários o valor apresentado é baixo quando comparado com os demais grupos.

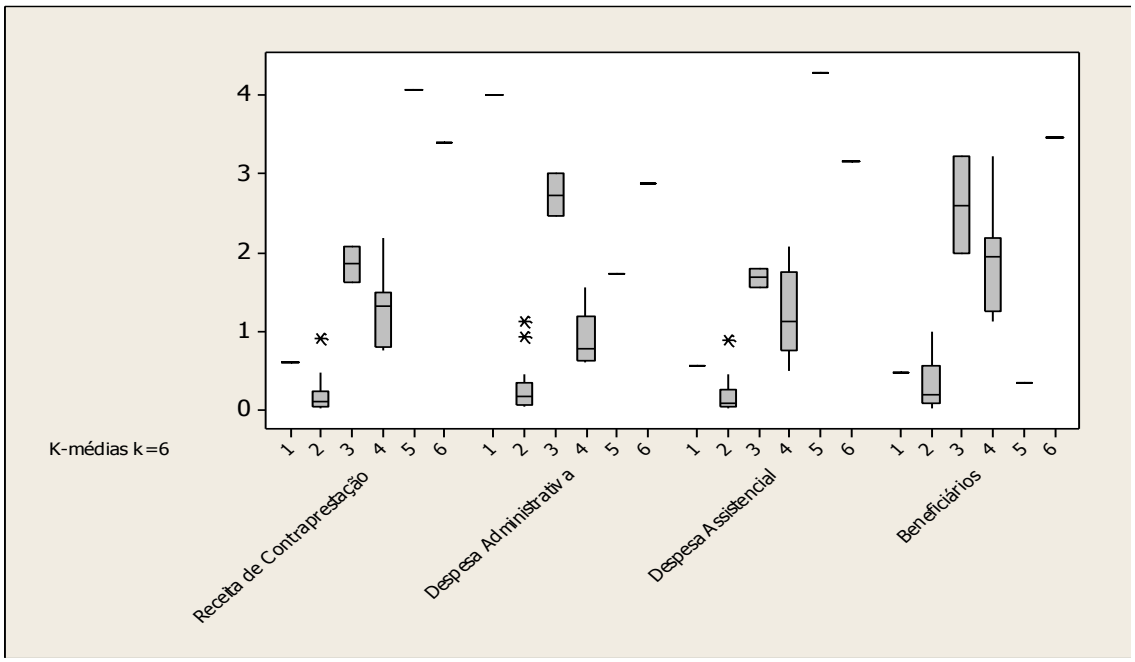
Grupo 3: Composto por duas operadoras, grupo caracterizado por possuir valores altos para todas as variáveis em estudo.

Grupo 4: Composto por seis operadoras e similar ao Grupo 3. Também se caracteriza por possuir valores altos para todas as variáveis, com exceção da variável despesa administrativa.

Grupo 5: Composto pela operadora 24, grupo caracterizado por possuir as maiores receitas de contraprestação e de despesa assistencial. Este grupo possui a menor média de beneficiários.

Grupo 6: Composto pela operadora 30, grupo caracterizado por possuir alta receita de contraprestação, despesa assistencial e a maior média de beneficiários.

Na Figura 9 observa-se o comportamento das variáveis para o agrupamento pelo método das K-médias para a composição de seis grupos.



Fonte: Base de dados em estudo

Figura 9 - Boxplot das variáveis padronizadas e agrupadas, K-médias k=6

Quadro 18 - Estatística Descritiva das variáveis padronizadas, K-Médias k=6

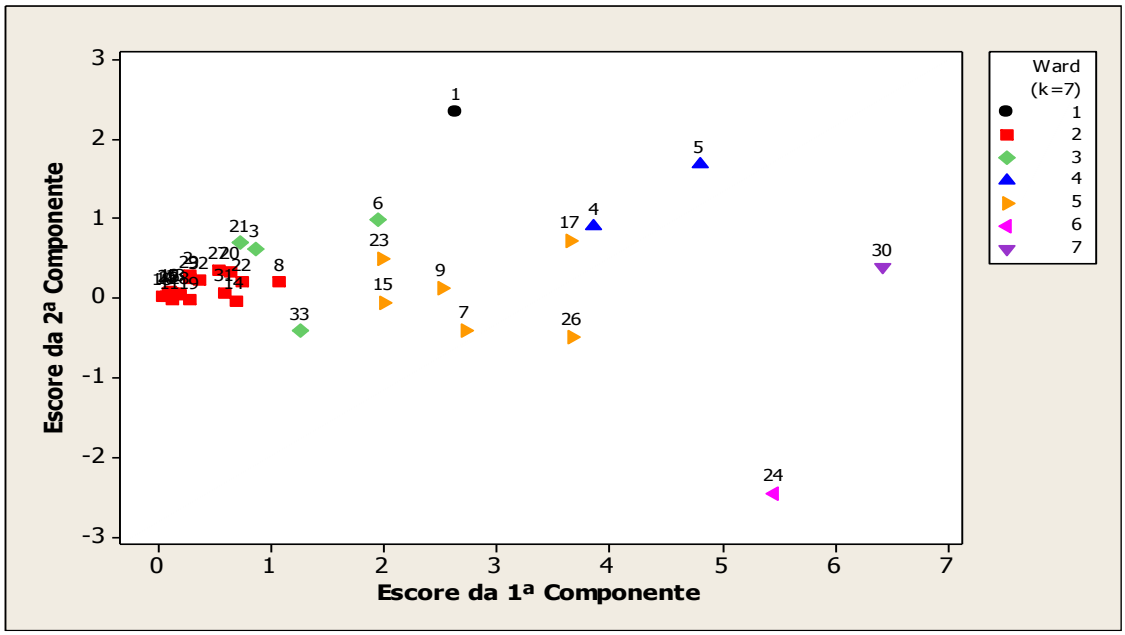
K-Médias k=6		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
nº de operadora		1	21	2	7	1	1
Receita de Contraprestação	Média	0,6060	0,1753	1,8505	1,3091	4,0771	3,3911
	Desvio-Padrão	-	0,2071	0,3302	0,4868	-	-
	Q_1	-	0,0354	-	0,7840	-	-
	Mediana	0,6060	0,1059	1,8510	1,3170	4,0771	3,3911
	Q_3	-	0,2242	-	1,4820	-	-
	Mínimo	0,6060	0,0140	1,6170	0,7479	4,0771	3,3911
	Máximo	0,6060	0,8974	2,0840	2,1860	4,0771	3,3911
Despesa Administrativa	Média	4,0136	0,2431	2,7324	0,8855	1,7180	2,8706
	Desvio-Padrão	-	0,2868	0,3753	0,3551	-	-
	Q_1	-	0,0482	-	0,6220	-	-
	Mediana	4,0136	0,1588	2,7320	0,7640	1,7180	2,8706
	Q_3	-	0,3313	-	1,1870	-	-
	Mínimo	4,0136	0,0263	2,4670	0,5876	1,7180	2,8706
	Máximo	4,0136	1,1104	2,9978	1,5533	1,7180	2,8706
Despesa Assistencial	Média	0,5567	0,1583	1,6795	1,2496	4,2943	3,1544
	Desvio-Padrão	-	0,2068	0,1729	0,5713	-	-
	Q_1	-	0,0248	-	0,7580	-	-
	Mediana	0,5567	0,0696	1,6795	1,1220	4,2943	3,1544
	Q_3	-	0,2523	-	1,7540	-	-
	Mínimo	0,5567	0,0056	1,5572	0,4799	4,2943	3,1544
	Máximo	0,5567	0,8712	1,8018	2,0731	4,2943	3,1544
Beneficiários	Média	0,4702	0,3279	2,5978	1,8644	0,3358	3,4588
	Desvio-Padrão	-	0,3049	0,8791	0,7272	-	-
	Q_1	-	0,0689	-	1,2450	-	-
	Mediana	0,4702	0,1833	2,5978	1,9490	0,3358	3,4588
	Q_3	-	0,5541	-	2,1870	-	-
	Mínimo	0,4702	0,0100	1,9762	1,1160	0,3358	3,4588
	Máximo	0,4702	0,9878	3,2193	3,2202	0,3358	3,4588

Fonte: Base de dados em estudo

A mudança da operadora 6 do Grupo 2 para o Grupo 4 no método das K-médias, torna o Grupo 2 o agrupamento com as menores médias para todas as variáveis. Essa alteração não modifica a interpretação do Grupo 4.

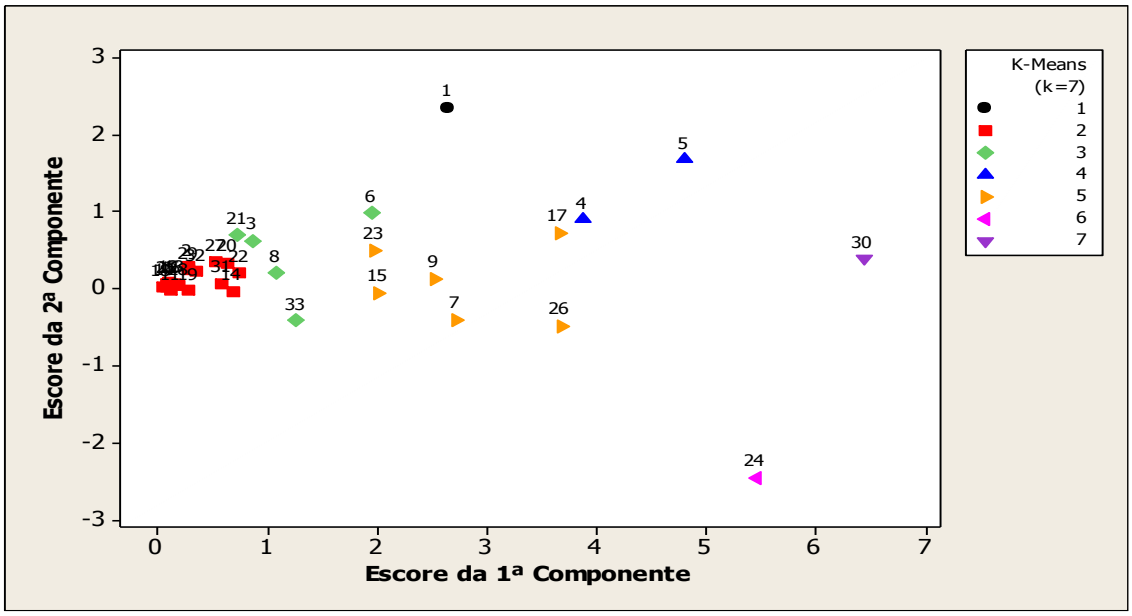
Comparação Ward e K-Médias para k=7 grupos

Nas Figura 10 e 11 tem-se os gráficos de dispersão dos dados identificando cada agrupamento formado. No método de Ward a operadora 8 é alocada no grupo dois e no das K-médias é alocada no grupo três, observa-se nos gráficos seguintes que esta operadora se encontra próxima dos dois grupos mencionados sendo difícil inferir em qual melhor se enquadra.



Fonte: Base de dados em estudo

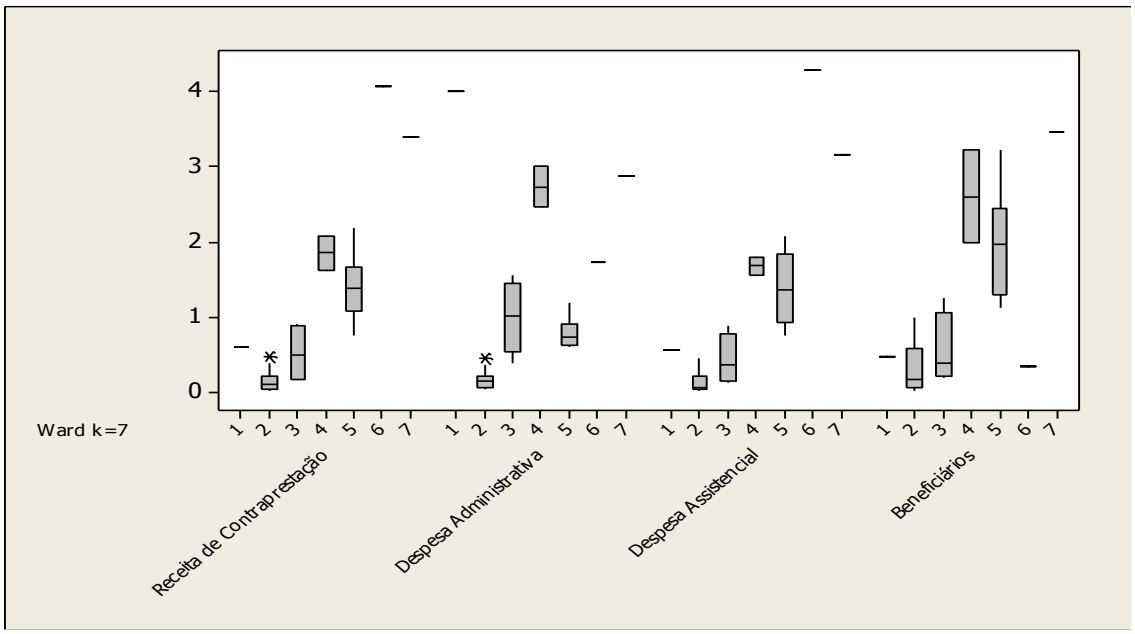
Figura 10 - Escore das duas componentes principais, Ward (k=7)



Fonte: Base de dados em estudo

Figura 11 - Escore das duas componentes principais, K-médias (k=7)

Na Figura 12 observa-se o comportamento das variáveis para o agrupamento pelo método Ward para a composição de sete grupos.



Fonte: Base de dados em estudo

Figura 12 - Boxplot das variáveis padronizadas e agrupadas, Ward k=7

Quadro 19 - Estatística Descritiva das variáveis padronizadas, Ward k=7

Ward k=7		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7
nº de operadora		1	18	4	2	6	1	1
Receita de Contraprestação	Média	0,6060	0,1363	0,5029	1,8505	1,3965	4,0771	3,3911
	Desvio-Padrão	-	0,1347	0,3930	0,3302	0,4691	-	-
	Q_1	-	0,0316	0,1600	-	1,0850	-	-
	Mediana	0,6060	0,0898	0,4790	1,8510	1,3830	4,0771	3,3911
	Q_3	-	0,1961	0,8690	-	1,6580	-	-
	Mínimo	0,6060	0,0140	0,1555	1,6170	0,7479	4,0771	3,3911
	Máximo	0,6060	0,4548	0,8974	2,0840	2,1860	4,0771	3,3911
Despesa Administrativa	Média	4,0136	0,1490	0,9941	2,7324	0,7742	1,7180	2,8706
	Desvio-Padrão	-	0,1203	0,4834	0,3753	0,2173	-	-
	Q_1	-	0,0438	0,5210	-	0,6133	-	-
	Mediana	4,0136	0,1313	1,0190	2,7320	0,7266	1,7180	2,8706
	Q_3	-	0,1963	1,4430	-	0,8936	-	-
	Mínimo	4,0136	0,0263	0,3856	2,4670	0,5876	1,7180	2,8706
	Máximo	4,0136	0,4396	1,5533	2,9978	1,1869	1,7180	2,8706
Despesa Assistencial	Média	0,5567	0,1161	0,4282	1,6795	1,3779	4,2943	3,1544
	Desvio-Padrão	-	0,1341	0,3313	0,1729	0,5033	-	-
	Q_1	-	0,0224	0,1490	-	0,9290	-	-
	Mediana	0,5567	0,0593	0,3620	1,6800	1,3480	4,2943	3,1544
	Q_3	-	0,2051	0,7730	-	1,8340	-	-
	Mínimo	0,5567	0,0056	0,1173	1,5572	0,7575	4,2943	3,1544
	Máximo	0,5567	0,4357	0,8712	1,8018	2,0731	4,2943	3,1544
Beneficiários	Média	0,4702	0,3299	0,5485	2,5978	1,9676	0,3358	3,4588
	Desvio-Padrão	-	0,3264	0,4810	0,8791	0,7383	-	-
	Q_1	-	0,0596	0,2080	-	1,2910	-	-
	Mediana	0,4702	0,1557	0,3830	2,5980	1,9670	0,3358	3,4588
	Q_3	-	0,5784	1,0550	-	2,4450	-	-
	Mínimo	0,4702	0,0100	0,1833	1,9762	1,1160	0,3358	3,4588
	Máximo	0,4702	0,9878	1,2451	3,2193	3,2202	0,3358	3,4588

Fonte: Base de dados em estudo

Analisando a Figura 8 e o Quadro 19 pode-se descrever o agrupamento formado:

Grupo 1: Composto somente pela operadora 1, grupo caracterizado por ter a maior média de gasto com despesa administrativa, possui valores baixos para as demais variáveis.

Grupo 2: Composto por dezoito operadoras, possui as menores médias para todas as variáveis em estudo.

Grupo 3: Composto por quatro operadoras, grupo caracterizado por possuir valores medianos para todas as variáveis em estudo.

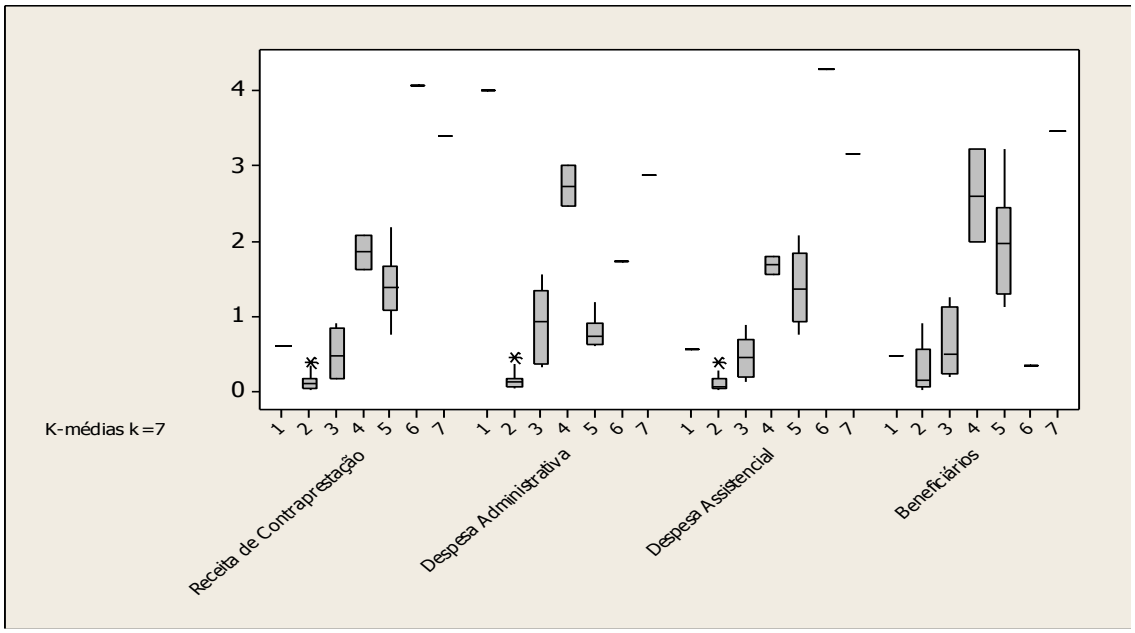
Grupo 4: Composto por duas operadoras, grupo caracterizado por possuir valores altos para todas as variáveis em estudo.

Grupo 5: Composto por seis operadoras, similar ao Grupo 4, porem com média baixa para a variável despesa administrativa.

Grupo 6: Composto pela operadora 24, grupo caracterizado por possuir a maior receita de contraprestação e de despesa assistencial. Este grupo possui a menor média de beneficiários.

Grupo 7: Composto pela operadora 30, grupo caracterizado por possuir alta receita de contraprestação, despesa assistencial e a maior média de beneficiários.

Na Figura 13 e Quadro 20 observa-se o comportamento das variáveis para o agrupamento pelo método das K-médias para a composição de sete grupos. Com a mudança da operadora 8 do Grupo 2 para o Grupo 3 não ocorre alteração na interpretação dos grupos.



Fonte: Base de dados em estudo

Figura 13- Boxplot das variáveis padronizadas e agrupadas, K-médias k=7

Quadro 20 - Estatística Descritiva das variáveis padronizadas, K-médias k=7

K-Médias k=7		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7
nº de operadora		1	18	4	2	6	1	1
Receita de Contraprestação	Média	0,6060	0,1176	0,4933	1,8505	1,3965	4,0771	3,3911
	Desvio-Padrão	-	0,1121	0,3410	0,3302	0,4691	-	-
	Q_1	-	0,0314	0,1650	-	1,0850	-	-
	Mediana	0,6060	0,0855	0,4550	1,8505	1,3830	4,0771	3,3911
	Q_3	-	0,1628	0,8410	-	1,6580	-	-
	Mínimo	0,6060	0,0140	0,1555	1,6170	0,7479	4,0771	3,3911
	Máximo	0,6060	0,3777	0,8974	2,0840	2,1860	4,0771	3,3911
Despesa Administrativa	Média	4,0136	0,1394	0,8577	2,7324	0,7742	1,7180	2,8706
	Desvio-Padrão	-	0,1167	0,5180	0,3753	0,2173	-	-
	Q_1	-	0,0436	0,3490	-	0,6133	-	-
	Mediana	4,0136	0,1208	0,9270	2,7324	0,7266	4,0771	3,3911
	Q_3	-	0,1695	1,3320	-	0,8936	-	-
	Mínimo	4,0136	0,0263	0,3120	2,4670	0,5876	1,7180	2,8706
	Máximo	4,0136	0,4396	1,5533	2,9978	1,1869	1,7180	2,8706
Despesa Assistencial	Média	0,5567	0,0973	0,4297	1,6795	1,3779	4,2943	3,1544
	Desvio-Padrão	-	0,1111	0,2870	0,1729	0,5033	-	-
	Q_1	-	0,0217	0,1810	-	0,9290	-	-
	Mediana	0,5567	0,0557	0,4360	1,6795	1,3480	4,2943	3,1544
	Q_3	-	0,1542	0,6760	-	1,8340	-	-
	Mínimo	0,5567	0,0056	0,1173	1,5572	0,7575	4,2943	3,1544
	Máximo	0,5567	0,3861	0,8712	1,8018	2,0731	4,2943	3,1544
Beneficiários	Média	0,4702	0,2912	0,6363	2,5978	1,9676	0,3358	3,4588
	Desvio-Padrão	-	0,2908	0,4605	0,8791	0,7383	-	-
	Q_1	-	0,0539	0,2320	-	1,2910	-	-
	Mediana	0,4702	0,1433	0,4840	2,5978	1,9670	0,3358	3,4588
	Q_3	-	0,5541	1,1160	-	2,4450	-	-
	Mínimo	0,4702	0,0100	0,1833	1,9762	1,1160	0,3358	3,4588
	Máximo	0,4702	0,8977	1,2451	3,2193	3,2202	0,3358	3,4588

Fonte: Base de dados em estudo

5 CONSIDERAÇÕES FINAIS

Sabe-se que a técnica estatística de análise de conglomerados auxilia na divisão de dados amostrais em grupos de acordo com a análise de similaridade presente nos elementos da amostra. Esta análise não é sujeita a uma só solução, pois o agrupamento escolhido pode se alterar conforme os critérios adotados para definir a qualidade da partição dos dados.

Neste estudo foi realizado inicialmente o agrupamento pelo método Ward com o número de grupos k igual 3, 4, 5, 6 e 7. A decisão sobre o melhor número de grupos foi realizada através da análise do nível de similaridade, do coeficiente de correlação intra-classe (R^2) e da estatística Pseudo F. Além disso, utilizou-se o método K-Médias, sendo as sementes iniciais os vetores de médias dos grupos formados pelo método de Ward.

A formação de três e quatro grupos foram descartadas, visto que o coeficiente de correlação intra-classe era inferior a 80%. A formação com cinco *clusters,* inicialmente apontada como valor mais adequado apresentou propostas de agrupamentos diferentes em cada método, inclusive no método K-médias o valor da estatística Pseudo F foi inferior ao apurado na formação de três e quatro grupos.

Diante do exposto, optou-se como melhores sugestões a formação de seis e sete *clusters.* Quando se realizou a análise destas duas formações via componentes principais foi possível visualizar as operadoras discrepantes no banco de dados (1, 4, 5, 24 e 30), elas possuem perfis bem distintos das demais. No caso das operadoras 1, 24 e 30 são responsáveis por modificar as características do grupo a que são alocadas e verificou-se que a qualidade do agrupamento melhora quando elas formam grupos “individuais”.

Um dos atributos da consultoria é o atendimento a inúmeros clientes o que causa impossibilidade de tratar problemas de forma mais específica. O grande volume de atendimento impede que esta seja mais individualizada, porém há um leque de possibilidades para melhorar e ampliar o atendimento para estes clientes. Visto que cada grupo apresenta suas características próprias, a consultoria pode trabalhar na criação de soluções para as principais deficiências apresentadas em cada *cluster* formado.

A consultoria atuarial tem por objetivo utilizar a análise de agrupamento exposta para montar uma equipe de atendimento personalizada para cada grupo de operadoras, dando um retorno mais ágil e eficiente ao cliente. Os resultados encontrados servirão

para auxiliar a elaboração de ações estratégicas relevantes ao mercado de saúde que melhorem o desempenho das operadoras em estudo.

O uso da análise de agrupamento traz inúmeras vantagens, pois é um método estatístico de fácil entendimento e inúmeras aplicações. Para o trabalho em questão, sua utilização conseguiu atender de forma satisfatória o objetivo de dividir a carteira de operadoras de planos de saúde da consultoria atuarial estudada em grupos com características semelhantes.

REFERÊNCIAS

ALMEIDA, C. O mercado privado de serviços de saúde: panorama atual e tendências da assistência médica suplementar. Brasília: IPEA, 1998, 80p.

BRASIL, AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR, A ANS: Quem Somos, Rio de Janeiro: ANS, 2011. Disponível em: <http://www.ans.gov.br/aans/quem-somos> Acesso em 27/04/2015.

BRASIL, AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR, Autorização de funcionamento das operadoras, Rio de Janeiro: ANS, 2014. Disponível em: <http://www.ans.gov.br/images/stories/Materiais_para_pesquisa/Materiais_por_assunto/guia_autorizacao_funcionamento_2014.pdf> Acesso em 10/05/2015.

BRASIL, MINISTÉRIO DA SAÚDE. Guia dos direitos do consumidor de seguros e planos de saúde, 1999. Disponível em: <http://www.planodesaudeoquesaber.com.br/public/data/guia/guia-fenasaude.pdf> Acesso em 10/05/2015.

CERVO, A.L.; BERVIAN, P. A. Metodologia científica. 5. ed. São Paulo: Prentice Hall, 2002, 242 p.

CORDEIRO, H. A. As empresas médicas: as transformações capitalistas da prática médica. 1. ed . Rio de Janeiro, Edições Graal, 1984, 175 p.

CORRAR, J. L. et al. Análise Multivariada para os cursos de administração, ciências contábeis e economia. 1. ed. São Paulo, Atlas, 2014, 541 p.

HAIR, J. F. BLACK W. C. BABIN B. J. ANDERSON R. E. TATHAM R. L. Análise multivariada de dados. 5. ed. Porto Alegre: Bookman, 2005, 593 p.

JONHSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. 6. ed. New Jersey: Prentice Hall, 2007, 607 p.

MESQUITA, J. M. C. Estatística multivariada aplicada à administração: guia prático para utilização do SPSS. Curitiba: Editora CRV, 2010, 167 p.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Ed. UFMG, 2005, 295 p.

NETO, J. M. M.; MOITA, G. Ciaramella. Uma introdução à análise exploratória de dados multivariado. 1997. Disponível em: <http://www.scielo.br/pdf/qn/v21n4/3193.pdf> Acesso em 16 novembro 2015.

TRION, R. C.,BAILEY, D.E. Cluster Analysis. New York: McGraw-Hill, 1970, 347 p.

ANEXO I

O Quadro A1 apresenta a seguir os valores observados das variáveis para cada uma das operadoras estudadas.

Quadro A1- Variáveis Originais

Operadora	Receita de Contraprestação	Despesa Administrativa	Despesa Assistencial	Beneficiários
1	24.921.775	40.502.169	19.330.756	9.876
2	2.147.569	4.435.860	1.346.406	1.372
3	7.175.175	9.354.536	8.491.619	10.164
4	66.499.325	24.895.350	62.565.904	41.507
5	85.706.703	30.251.809	54.073.174	67.619
6	32.253.312	15.674.887	16.663.671	26.151
7	60.932.089	7.711.277	60.911.856	28.339
8	18.701.935	3.148.691	15.127.729	20.748
9	54.144.237	6.275.916	38.965.645	41.685
10	1.313.454	343.798	801.600	1.524
11	3.514.946	528.817	2.415.292	457
12	1.599.302	1.219.006	557.199	893
13	2.952.851	1.638.986	1.933.624	663
14	15.531.502	2.576.694	13.407.848	7.158
15	49.224.255	6.953.462	34.255.849	23.441
16	1.265.455	435.676	921.378	3.010
17	59.643.233	11.977.413	54.653.625	67.637
18	632.431	444.341	702.138	209
19	6.391.168	1.430.160	6.486.583	1.740
20	6.997.786	1.602.291	4.221.814	18.856
21	6.395.582	11.205.680	4.071.978	3.849
22	13.671.005	3.537.879	9.209.293	12.366
23	30.755.979	5.930.120	26.303.583	40.929
24	167.671.337	17.336.579	149.115.380	7.054
25	1.057.016	426.672	234.493	2.462
26	89.899.799	8.030.388	71.985.555	45.937
27	5.051.367	1.782.750	2.187.397	16.022
28	4.353.566	649.865	2.434.069	3.531
29	576.115	265.251	195.336	12.075
30	139.459.869	28.968.109	109.532.153	72.649
31	11.262.962	1.616.194	9.026.922	10.432
32	3.871.985	985.866	1.379.349	11.202
33	36.907.486	3.891.565	30.250.947	5.916

No Quadro A2 tem-se os valores observados das variáveis padronizadas para cada uma das operadoras estudadas.

Quadro A2 - Variáveis após a Padronização

Operadora	Receita de Contraprestação	Despesa Administrativa	Despesa Assistencial	Beneficiários
1	0,6060	4,0136	0,5567	0,4702
2	0,0522	0,4396	0,0388	0,0653
3	0,1745	0,9270	0,2445	0,4839
4	1,6170	2,4670	1,8018	1,9762
5	2,0840	2,9978	1,5572	3,2193
6	0,7843	1,5533	0,4799	1,2451
7	1,4816	0,7642	1,7542	1,3492
8	0,4548	0,3120	0,4357	0,9878
9	1,3166	0,6219	1,1222	1,9846
10	0,0319	0,0341	0,0231	0,0726
11	0,0855	0,0524	0,0696	0,0218
12	0,0389	0,1208	0,0160	0,0425
13	0,0718	0,1624	0,0557	0,0316
14	0,3777	0,2553	0,3861	0,3408
15	1,1969	0,6891	0,9865	1,1160
16	0,0308	0,0432	0,0265	0,1433
17	1,4503	1,1869	1,5740	3,2202
18	0,0154	0,0440	0,0202	0,0100
19	0,1554	0,1417	0,1868	0,0828
20	0,1702	0,1588	0,1216	0,8977
21	0,1555	1,1104	0,1173	0,1833
22	0,3324	0,3506	0,2652	0,5887
23	0,7479	0,5876	0,7575	1,9486
24	4,0771	1,7180	4,2943	0,3358
25	0,0257	0,0423	0,0068	0,1172
26	2,1860	0,7958	2,0731	2,1871
27	0,1228	0,1767	0,0630	0,7628
28	0,1059	0,0644	0,0701	0,1681
29	0,0140	0,0263	0,0056	0,5749
30	3,3911	2,8706	3,1544	3,4588
31	0,2739	0,1602	0,2600	0,4967
32	0,0942	0,0977	0,0397	0,5333
33	0,8974	0,3856	0,8712	0,2817