Paola Mara de Oliveira Quinto

Análise de cluster em um plano de saúde via wavelets

Belo Horizonte, fevereiro de 2013

Paola Mara de Oliveira Quinto

Análise de cluster em um plano de saúde via wavelets

Dissertação apresentada como requisito parcial
para obtenção de grau de Mestre em Estatística
pela Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Renato Martins Assunção

Programa de Pós-Graduação em Estatística
Departamento de Estatística
Instituto de Ciências Exatas
Universidade Federal de Minas Gerais

Belo Horizonte, fevereiro de 2013

Agradecimentos

Agradeço a Deus por ter me ajudado, me dado força e entendimento durante essa etapa da vida.

Aos meus pais Deocleciano e Gislaene, pelo amor, carinho, compreensão e por sempre me incentivarem. Enfrentamos juntos a saudade, mas mesmo distantes fisicamente, as orações e o apoio de vocês sempre foi e será essencial na minha vida. Ao meu irmão Diego, que sempre esteve presente, ouvindo minha reclamações, me aguentou nos momentos de nervosismo, obrigada pela paciência e compreensão.

Ao meu orientador, Professor Renato Assunção, pelo apoio, pela paciência e compreensão. Pelas explicações sempre objetivas e claras e por ter aceitado ser meu orientador.

Aos professores do curso de mestrado em estatística, pelo conhecimento transmitido.

Aos membros da banca examinadora, Ana Paula Viegas (UNIMED-BH), Prof.Fábio Demarqui(UFMG) e Prof. Wagner Barreto (USP), pela leitura, correções e sugestões da dissertação, que foram essenciais para o meu crescimento e término deste trabalho. Agradeço a Ana Paula, que na função de chefe, permitiu a aplicação do método estudado na operadora de planos de saúde UNIMED-BH. Obrigada pela confiança e pelos conhecimentos obtidos, os quais ampliaram minha visão acerca dos planos de saúde.

Aos meus pastores Chicão e Priscila por sempre estarem ao meu lado, pelos conselhos e orações. Às minhas amigas Rosana, Fernanda e Márcia que sempre me ajudaram. Muito obrigada!

À CAPES pela bolsa de mestrado.

Resumo

Um dos assuntos que tem trazido preocupações por parte das operadoras de planos de saúde, e ao mesmo tempo, tem sido alvo de muitos estudos, é a elevação dos custos e sua concentração em uma pequena parcela da carteira. Dentro deste contexto, diferentes tipos de clientes são responsáveis por compor os planos e gerar os custos. No entanto, não se sabe quantos existem e quais as características peculiares de cada um deles, e nosso objetivo neste trabalho será identificá-los. A base de dados utilizada é de um determinado plano de saúde, e o método adotado para separar os clientes dessa carteira em grupos ou perfis é denominado análise de cluster. O propósito da análise de cluster é buscar uma classificação de acordo com as relações naturais que a amostra apresenta, formando grupos de objetos por similaridade. Porém, quando aplicado à base de dados em questão, o método não consegue separar os clientes em grupos com características homogêneas de acordo com os custos. Buscamos, então, uma forma de reescrevê-los através dos coeficientes de wavelets, os quais resumem toda a informação contida nas séries históricas dos custos de cada cliente do plano de saúde. Várias análises foram realizadas, mas traremos a que obteve melhores resultados. Descreveremos os perfis de clientes formados, assim como suas características com relação às séries dos custos e às descritivas gerais do grupo, tais como idade, sexo, custo total, titularidade, entre outras.

Palavras-chaves: *Análise de cluster, método k-means, wavelets, análise de resolução múltiplas, perfis de clientes.*

Abstract

One of the issues that has led to concerns by operators of health plans, and at the same time, has been the subject of many studies, is rising costs and their concentration in a small portion of the portfolio. Within this context, different types of customers are responsible for writing plans and generate costs. However, no one knows how many there are and what the characteristics of each one of them, and our goal in this work is to identify them. The database is used for a particular health plan, and the method used to separate customers into groups is called cluster analysis. The purpose of cluster analysis is to seek a classification according to the natural features that the sample, forming groups of objects by similarity. However, when applied to the database in question, the method fails to separate customers in groups with homogeneous characteristics according costs. Thean, we look a way to rewrite the costs through the wavelet coefficients, which summarize all the information contained in the time series of the costs of each client's health plan. Several analysis were performed, but we will bring the better result. We describe the customer profiles formed, as well as their characteristics with respect to the series of costs and descriptive general group, such as age, sex, total cost ownership, among others.

Keywords: *Cluster analysis, k-means method, wavelets, multiresolution analysis, customer profiling.*

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Ao longo dos últimos anos, o número de beneficiários de planos de saúde cresceu consideravelmente devido à estabilidade e ao crescimento econômico, os quais permitiram a elevação dos indicadores de emprego e de renda dos trabalhadores brasileiros que, consequentemente, passaram a gastar frações maiores dos salários com saúde. Tal crescimento apresenta uma distinção marcante quanto ao tipo de contratação, que pode ser individual/familiar ou coletiva, esse último com participação e crescimento mais significativos que o primeiro. A contratação coletiva, em geral, está relacionada ao mercado de trabalho pelo fato de ser o segmento no qual o contratante são pessoas jurídicas, diferente do mercado individual, no qual o contratante são pessoas físicas (Leal e Matos, 2007). Ressaltamos que, no momento, os planos empresariais tem rejuvenescido a carteira dos planos.

Dentre as principais discussões na aérea de saúde suplementar, destaca-se a importância da avaliação dos custos, pois estes vêm crescendo progressivamente. Vários aspectos corroboram para esse aumento: incremento de novas tecnologias médicas, aumento do uso de exames, aumento da longevidade, diminuição da taxa de fecundidade, aumento da renda, transição epidemiológica, entre outros. As maiores exigências do órgão regulador sobre as garantias financeiras e sobre o rol de procedimentos, também têm impactado a já apertada margem de lucro das operadoras. Todo este cenário afeta a sustentabilidade econômico-financeira das operadoras, ou seja, o equilíbrio intemporal de suas contas e sua estruturação econômica de forma a suportar as despesas demandadas no longo prazo, uma vez que os custos não tendem a reduzir nos próximos anos, pelo contrário, tendem a aumentar cada vez mais.

A expansão do sistema de saúde suplementar nas últimas décadas foi significativa, estimando-se que, atualmente, cerca de um quarto da população está associada a algum tipo de plano, conforme informações da Agência Nacional de Saúde Suplementar (ANS, 2012). Para manter sustentável o setor, a ANS preconiza uma sinistralidade de 70%. No entanto, segundo ela, a sinistralidade girou em torno de 82% em 2011.

Isso revela que a receita tende a não acompanhar o crescimento dos custos, principalmente quando tratamos de planos individuais, nos quais os reajustes dos prêmios são definidos pela ANS. Quanto aos planos coletivos, os reajustes são definidos pelo equilíbrio contratual, sem intervenções do órgão regulador do sistema de saúde, sendo que os reajustes podem chegar em um ponto que a empresa não consiga manter o contrato com a operadora e acabe buscando um

preço menor com a concorrência. Segundo Lima e Lima (1998), além dos custos crescentes, um dos principais problemas das organizações de saúde é a ineficiência. Por essa razão, torna-se essencial o aprimoramento da administração dos custos e eficiência na prestação de serviços de saúde (Medici e Marques, 1996). Além disso, na esteira da temática dos custos, temos visto que uma pequena porcentagem dos beneficiários é, de fato, responsável pela maior porcentagem dos custos totais em uma empresa ou operadora de planos de saúde. Segundo Ailon et al.(2005), cerca de 20% dos clientes em uma operadora é responsável por, aproximadamente, 80% dos custos totais médicos anuais, o que mostra um caráter aleatório e altamente concentrado dos custos.

No entanto, os clientes que são responsáveis por gerar elevados custos hoje, não serão necessariamente responsáveis por acarretá-los no futuro. Por isso, muitas pesquisas têm sido realizadas a fim de encontrar modelos que predigam quem serão os indivíduos de alto custo: aqueles que são responsáveis por realizar gastos dispendiosos e que, por sua vez, consomem grande parte dos custos totais em uma operadora. E ainda, suponha que um cliente não tenha gerado nenhum custo em vários meses, almejamos saber a probabilidade dele obter algum custo ou nenhum custo nos próximos meses. Tanto na saúde pública, quanto na suplementar, desenvolver metodologias que permitam identificar grupos populacionais de alto custo ou predizer futuros grupos de alto custo é necessário para a sustentabilidade do setor.

Quando tentamos entender o que leva a esta distorção e concentração dos gastos ou custos nos planos de saúde, encontramos muitos usuários que utilizam os planos de forma inadequada. Por exemplo, encontramos usuários realmente doentes, e que por isso necessitam de tratamento, mas não seguem apenas um determinado médico, fazem repetidos exames diagnósticos, realizam procedimentos sem uma orientação única, o que não resolve seu problema de saúde e gera custos elevados. Conjuntamente, temos usuários que realizam procedimentos médicos de forma esporádica e que geram baixos custos, como também aqueles que realizam procedimentos constantes sem realmente estarem doentes e produzem custos para as operadoras.

Portanto, vemos que diferentes perfis ou tipos de clientes compõem as carteiras dos planos, no que diz respeito ao comportamento das séries históricas dos custos de cada um deles. Porém, não temos conhecimento de quantas e quais tipologias existem, e almejamos identificá-las e entendê-las. Métodos de análise de cluster serão utilizados para identificar os grupos de clientes semelhantes quanto ao perfil de despesas, e cada grupo resultante da análise de cluster representará um perfil a ser estudado.

Essa é a nossa principal contribuição, uma vez que um método aplicável aos dados de uma operadora e que permita a gestão das despesas assistenciais da carteira de clientes é fundamental para desenvolvimento de ações de gestão da saúde segmentado por grupos populacionais e por linhas de cuidado, que mitiguem os custos. Somente assim, será possível garantir acesso aos serviços de saúde com qualidade para a população de beneficiários que tende ficar mais velha, longeva e com um estado de morbidade que requer cuidados específicos. E, quanto aos tipos de clientes que acarretam custos abusivos e desordenados, as operadoras podem agir com ações educativas e assistenciais que alterem hábitos e melhorem a qualidade de saúde dos mesmos.

Em suma, o objetivo geral deste trabalho é avaliar a aplicabilidade do método de análise de cluster associado às wavelets para criar tipologias de clientes a partir das despesas assistenciais apresentadas em agosto de 2003 a novembro de 2008 por uma carteira de 99.865 clientes de uma operadora de saúde e identificar quantos tipos de clientes existem na carteira desse plano de saúde, baseando-se nas series históricas de despesas assistenciais e descrever as características particulares de cada um deles.

O texto a seguir está organizado da seguinte forma: o próximo capítulo abordará conceitos e definições sobre análise de cluster, assim como alguns métodos existentes, atentando-se para o método K-médias. O Capítulo 3 abordará os principais conceitos sobre as wavelets, além dos algoritmos para decomposição e para a reconstrução das funções através dos coeficientes de wavelets. Sendo assim, os Capítulos 2 e 3 constituirão o referencial teórico deste trabalho. Na Seção 2.5 do Capítulo 2, explicaremos porque, sem as wavelets, os métodos de análise de cluster não conseguem identificar os grupos de clientes. No Capítulo 4, traremos alguns procedimentos metodológicos. No Capítulo 5 apresentaremos os resultados, e por fim, no Capítulo 6, teremos as conclusões e algumas considerações finais.

2 Análise de cluster ou agrupamentos

O termo análise de cluster ou análise de agrupamentos, primeiramente usado por Tyron (1939), é um conjunto de técnicas estatísticas cujo objetivo é separar os elementos da amostra em grupos ou conglomerados homogêneos, de forma que cada partição ou grupo seja similar com respeito a algum critério ou característica. Os elementos em cada conglomerado tendem a ser semelhantes entre si, porém diferentes dos demais elementos em outros conglomerados. Cada grupo obtido deve apresentar tanto uma homogeneidade interna (dentro de cada grupo), como uma grande heterogeneidade externa (entre grupos).

Várias são as situações onde a análise de agrupamentos se faz presente: em pesquisas de mercado, na segmentação de clientes de acordo com perfis de consumo; em Ecologia na classificação de espécies; em Geografia, na classificação de cidades, estados, etc; na classificação de pessoas de acordo com seus perfis de personalidade.

Na figura abaixo temos um exemplo de agrupamento: cada sinal + corresponde a um indivíduo, sendo que aqueles que se encontram em uma mesma região delimitada (grupo) são similares de acordo com as doenças X e Z. O grupo cujos elementos são circundados é composto pelos indivíduos similares com relação à doença X, enquanto o grupo cujos elementos não são circundados é composto pelos indivíduos similares com relação à doença Z.

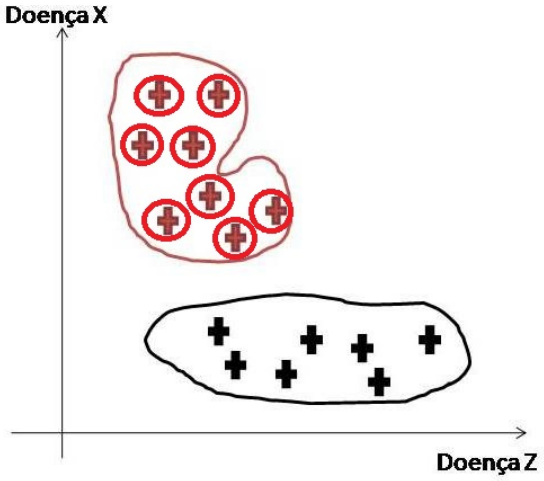


Figura 1: Exemplos de agrupamentos

Nas seções seguintes, apresentaremos os critérios de escolha de similaridade e dissimilaridade entre os elementos e o algoritmo de formação dos agrupamentos.

2.1 Medidas de Similaridade e Dissimilaridade

Suponha que temos n elementos amostrais, tendo-se medido p variáveis (em cada um deles) a serem utilizadas como critério de semelhança entre eles. Para cada elemento amostral $j \in \{1, 2...n\}$ denotamos:

$$X_j = [X_{1j}, X_{2j}, ...X_{pj}]$$

Um conceito fundamental na utilização das técnicas de análise de agrupamento é a escolha de um critério (ou medida) que meça a distância entre dois objetos, ou que quantifique o quanto eles são parecidos. Cabe observar que, tecnicamente, é possível dividir essa medida em duas categorias: medidas de similaridade e de dissimilaridade.

Na medida de similaridade, quanto maiores os valores observados, mais parecidos serão os objetos. Já para a medida de dissimilaridade, quanto maiores os valores observados, menos parecidos (mais dissimilares) serão os objetos. Existem várias medidas diferentes e cada uma delas produz um tipo de agrupamento.

A maioria dos algoritmos de análise de cluster estão programados para operarem com o conceito de distância (dissimilaridade). Os objetos com menor distância entre si são mais semelhantes, logo são aglomerados em um mesmo cluster. Já os mais distantes participam de clusters (conglomerados) distintos. Dentre as várias formas de medir a distância entre os objetos, a mais utilizada é a distância euclidiana, a qual será utilizada neste estudo. A distância euclidiana é calculada como a raiz quadrada da soma dos quadrados das diferenças de valores para cada variável. Temos também a distância de Mahalanobis, a distância de Minkowsky, entre outras.

Definição 2.1.1. *Distância Euclidiana: é a distância entre dois elementos amostrais X_l e X_k, baseada nos p atributos que os compõem:*

$$d(X_l, X_k) = [(X_l - X_k)'(X_l - X_k)]^{1/2} = [\sum_{i=1}^{p}(X_{il} - X_{ik})^2]^{1/2} \quad (1)$$

A TAB.1 fornece o custo mensal de três clientes de um plano de saúde. Eles foram criados apenas para facilitar nosso entendimento. Posteriormente, temos a distância euclidiana entre eles. O vetor aleatório $[X_{1j}, X_{2j}, ...X_{pj}]$ de cada cliente associado ao j, para $j = (1, 2, 3)$, é composto por $p = 7$ custos mensais. Portanto, a distância para cada cliente é calculada como a soma das distâncias entre os custos calculados mês a mês.

A distância entre os indivíduos 1 e 2 será:

Tabela 1: Dados artificiais

∗	Mês1	Mês2	Mês3	Mês4	Mês5	Mês6	Mês7
Cliente 1	0	250	0	0	0	0	0
Cliente 2	0	0	0	300	0	0	0
Cliente 3	0	100	30	50	0	15	0

$$d(X_1, X_2) = ((0-0)^2 + (250-0)^2 + (0-0)^2 + (0-300)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2)^{1/2} = \sqrt{152500} = 390,52$$

A distância entre os indivíduos 1 e 3 será:

$$d(X_1, X_3) = ((0-0)^2 + (250-100)^2 + (0-30)^2 + (0-50)^2 + (0-0)^2 + (0-15)^2 + (0-0)^2)^{1/2} = \sqrt{26125} = 161,63$$

Quanto maior a distância euclidiana, menos parecidos são os clientes. Assim, pelos cálculos feitos acima, o cliente 1 é mais similar ao cliente 3 e menos similar ao cliente 2.

2.2 Técnicas para a construção dos clusters

As técnicas de clusters são frequentemente classificadas em dois tipos: técnicas hierárquicas e não hierárquicas. As primeiras são classificadas em aglomerativas e divisivas e têm como objetivos identificar os possíveis grupos existentes e o valor provável do número de grupos. Os métodos de agrupamentos hierárquicos aglomerativos mais comuns e disponíveis na grande maioria dos *softwares* estatísticos são: método de ligação simples, método de ligação completa, método de Ward, método da média das distâncias, entre outros. Eles partem de uma matriz de distância ou similaridade entre os elementos da amostra.

Os gráficos denominados dendogramas podem ser construídos nesses casos. Esses fornecem o histórico dos agrupamentos: a escala vertical indica o nível de similaridade (ou dissimilaridade) e a escala horizontal indica os elementos amostrais numa ordem relacionada à história do agrupamento. É importante ressaltar que, uma vez unidos, os elementos amostrais não poderão ser separados.

Para o uso das técnicas não hierárquicas é necessário definir *a priori* o número de grupos, de forma que a partição satisfaça dois requisitos básicos: "coesão" interna ("semelhança" interna), e isolamento (ou separação) dos clusters formados.

A cada passo do algoritmo, novos grupos podem ser formados através da junção ou divisão de grupos criados em passos anteriores. Isto é, indivíduos colocados num mesmo conglomerado

em algum passo do algoritmo, não necessariamente "estarão juntos" no final da partição. Por isso, não é possível construir dendogramas. Os métodos k-médias *(K-means)* e Fuzzy c-médias *(Fuzzy c-means)* são os mais utilizados dentre as técnicas não hierárquicas.

As técnicas não hierárquicas em comparação às técnicas hierárquicas possuem maior eficiência ao tratar grandes conjuntos de dados, pois a matriz de distâncias não precisa ser determinada. No entanto, temos como desvantagem a especificação inicial do número de clusters k, a não ser que o pesquisador tenha um conhecimento *a priori* desse número. Ambas as técnicas de análise de cluster são sensíveis à ruídos, ou seja, observações com valores altos podem causar uma grande alteração nos resultados.

Observação: utilizaremos os termos clusters, conglomerados e agrupamentos para designar os grupos formados na análise de cluster.

2.3 Método k-means

Neste trabalho atentaremos para o método *k-means*, desenvolvido por Stuart Lloyd, em 1957. É, provavelmente, um dos métodos mais conhecidos e utilizados em problemas práticos. Lembramos que, como premissa dos métodos não hierárquicos, o número de grupos ou clusters k deve ser especificado. Temos a seguir a idéia geral do método k-means:

1. Escolhe-se arbitrariamente k objetos (sementes iniciais) $p_1....p_k$ do banco de dados. Estes objetos serão os centróides de k clusters, cada cluster D_i formado somente pelo objeto p_i, para $i = 1, ...k$. Os centróides representam a média das variáveis as quais caracterizam os indivíduos de cada grupo.

2. Cada elemento O_j do conjunto de dados, $(j = 1, ... n)$, em que n é o tamanho amostral, é então comparado com cada centróide inicial p_i, através de alguma medida de distância.

3. O elemento é alocado ao grupo cuja distância é a menor. Ou seja, passa a integrar o cluster representado por p_i.

4. Calcula-se a média dos elementos de cada cluster. Este ponto será o novo representante do cluster.

Em seguida, volta-se para o passo 2 : varre-se o banco de dados inteiro e para cada objeto O_j calcula-se a distância entre este objeto O_j e os novos centros. O objeto O_j será realocado

para o cluster D_i tal que **a distância entre O_j e o centro de D_i é a menor possível.**

Quando todos os objetos forem devidamente realocados entre os clusters, calcula-se os **novos centros dos clusters**. O processo se repete até que nenhuma realocação de elementos seja necessária ou até que os centróides não se alterem substancialmente. A figura abaixo ilustra o funcionamento do método k-means para $k = 3$. Na primeira iteração, os objetos circundados foram escolhidos aleatoriamente. Nas próximas iterações, os centróides são marcados com o sinal +. Vemos que a cada passo do algoritmo alguns objetos mudam de grupos. Ou seja, objetos unidos no primeiro passo do algoritmo, não estarão necessariamente juntos nos segundo e terceiro passos.

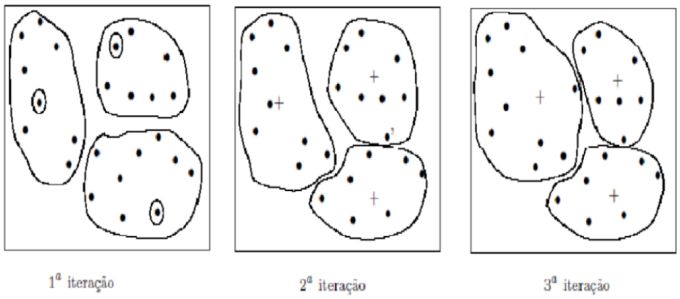


Figura 2: Exemplo do método k-means para k=3

Ressaltamos que a escolha das sementes iniciais de agrupamento influencia no agrupamento final e por isso, alguns cuidados precisam ser tomados. A maioria dos *softwares* estatísticos usa como *default* as k primeiras observações do banco de dados. O método pode trazer bons resultados quando esses elementos amostrais escolhidos inicialmente são discrepantes entre si, mas não é recomendável quando os elementos são semelhantes entre si. No *software* R, o qual utilizamos para implementar o k-means, as características de cada elemento amostral devem ser armazenadas em uma matriz. Elementos escolhidos aleatoriamente de diferentes linhas dessa matriz são selecionados como centróides iniciais.

O método k-means busca minimizar a soma dos erros, dada por:

$$\sum_{i=1}^{k} \sum_{x \in D_i} d(x, p_i),$$

onde $d(x, p_i)$ representa a distância do elemento ou objeto amostral x até o centróide p_i do cluster D_i. O algoritmo termina quando o erro não mais decresce significativamente, ou seja, quando não há mais troca dos elementos entre grupos.

2.4 Critério para escolha do número de clusters

Uma questão de grande importância na análise de agrupamentos via métodos não hierárquicos é a escolha do número de grupos k, que definem a partição de um conjunto de dados. Busca-se a melhor partição de ordem k, através de algum critério que forneça sua qualidade.

Computacionalmente, é inviável criar todas as partições possíveis de ordem k para um mesmo conjunto de dados, a não ser que o número de objetos seja bastante pequeno.

Como critério de escolha do número de clustes, adotaremos neste trabalho uma medida, a qual denominamos **medida de homogeneidade**, definida como:

$$SQD/SQE \tag{2}$$

onde SQD representa a Soma de Quadrados dentro dos grupos e SQE representa a Soma de Quadrados entre os grupos.

Seja $X'_{ij} = (X_{i1}, X_{i2}...X_{ip})$, o vetor com p medidas observadas para o j-ésimo elemento amostral do i-ésimo grupo; $\overline{X}^{(i)\prime} = (\overline{X_1^{(i)}}, \overline{X_2^{(i)}}...\overline{X_p^{(i)}})$, o vetor de médias do i-ésimo grupo; $\overline{X}' = (\overline{X_1}, \overline{X_2}...\overline{X_p})$, o vetor de médias global e n_i o número de elementos amostrais em cada cluster i.

As somas de quadrados entre e dentro dos grupos são definidas por:

$$SQD = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}^{(i)})'(X_{ij} - \overline{X}^{(i)}) \tag{3}$$

$$SQE = \sum_{i=1}^{k} n_i(\overline{X}^{(i)} - \overline{X})'(\overline{X}^{(i)} - \overline{X}) \tag{4}$$

A soma de quadrados entre os grupos representa a variabilidade entre os grupos, enquanto a soma de quadrados dentro dos grupos representa a variabilidade em cada um dos grupos. Queremos uma possível partição que forneça a menor variabilidade dentro dos grupos e a maior variabilidade entre os grupos. À medida que aumentamos o número de grupos, a medida de homogeneidade (2) decai, pois quanto maior o número de grupos, maior será a variabilidade entre os grupos (denominador) e menor será a variabilidade dentro dos grupos.

Podemos adotar o seguinte critério: escolhemos um número de clusters tal que, a partir dele, a medida de homogeneidade referente às numerações posteriores não decaiam de forma acentuada. Ou ainda, as medidas subsequentes sejam relativamente próximas umas das outras.

2.5 Tipologia de clientes

Baseando-se na série histórica dos custos gerados por cada cliente em um plano de saúde, nosso foco está em separá-los em grupos com características (ou padrões) similares entre si, no que se refere ao comportamento dos custos mensais e totais. Cada grupo representará um perfil de cliente. Outros atributos também podem ser considerados tais como sexo, idade, tipo de contratação, tipo de produto, titularidade, entre outros.

Para um melhor entendimento, suponha que dois clientes obtiveram custo total anual de R$15.000 gastos diferentemente: o primeiro cliente não teve custos em onze meses, ao passo que em um mês qualquer seu gasto foi de 15.000 reais. No caso do segundo cliente, o custo total foi diluído durante os doze meses: em alguns meses o custo foi zero e em outros meses, o custo esteve entre 500 e 3.000 reais. Veja a FIG.3. Observe que, embora ambos os clientes tenham o mesmo custo total anual, o perfil dos gastos é distinto e por isso, eles deveriam ser alocados em diferentes grupos que representam diferentes perfis quanto à série dos custos.

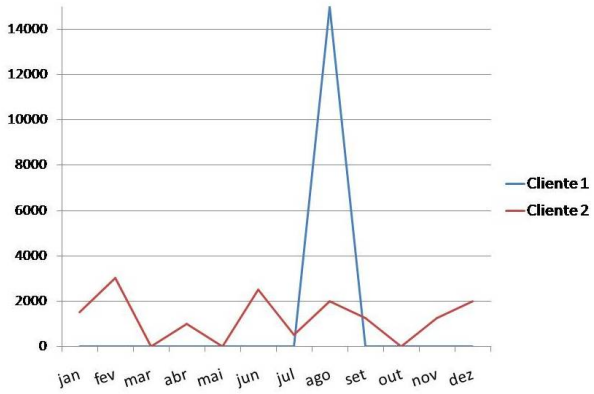


Figura 3: Série temporal do custo mensal de cada cliente

As Figuras 4 e 5 mostram dois exemplos adicionais: o primeiro gráfico representa um único perfil de cliente, assim como o segundo gráfico representa um outro perfil.

A FIG. 4 representa um perfil ou tipo de cliente que obtém custos maiores que zero em quase todos os meses da análise (de janeiro de 2006 a julho de 2007). Muitos meses com picos superiores a R$ 70 são observados, embora o custo total não é extremamente elevado - 860 reais para o primeiro cliente (gráfico à esquerda) e 1.000 reais para o segundo cliente (gráfico à direita). A operadora poderia intervir neste perfil com ações de assistência e prevenção de doenças, pois podem ser clientes realmente doentes ou clientes que usam o plano de forma desordenada. Ressalta-se que picos correspondem aos meses com custo elevado ou com custo

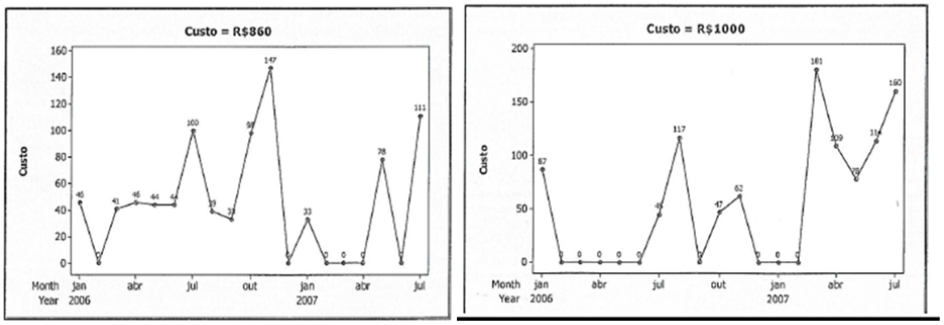


Figura 4: Exemplo I - Tipologia de cliente com base nos custos

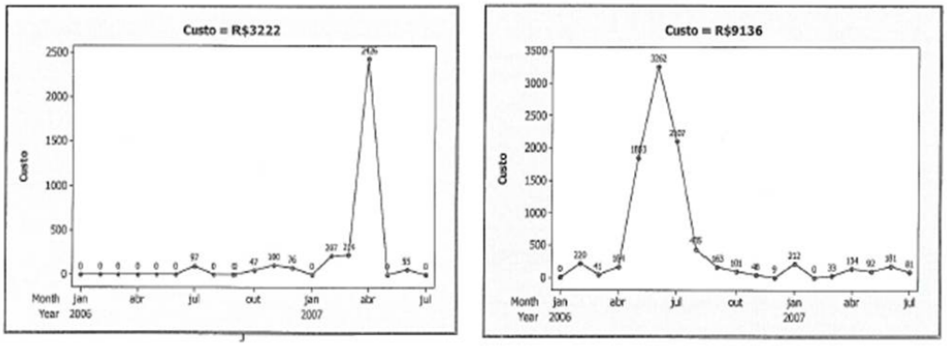


Figura 5: Exemplo II - Tipologia de cliente com base nos custos

superior ao custo médio da série histórica.

A FIG. 5 representa um perfil de cliente que obtém custo zero ou próximo de zero em quase todos os meses, e apenas em um mês tem custo superior a R$ 2.000, e esse valor é próximo ao custo total obtido nos dezenove meses. Tal perfil corresponde ao clientes que eventualmente utilizam o plano e não precisariam de intervenção imediata por parte da operadora já que, provavelmente, não são clientes que se encontram doentes ou realizam procedimentos médicos sem necessidade e desordenadamente. O custo total dos clientes desse perfil são 3.222 reais para o primeiro (gráfico à esquerda) e 9.136 reais para o segundo(gráfico à direita).

Também observamos que o custo total de ambos os clientes da FIG.5 é superior ao custo total dos clientes da FIG.4. Portanto, devido a todas essas descrições, as figuras acima correspondem a diferentes perfis de clientes que compõem uma carteira de planos de saúde.

Na visão da operadora de planos de saúde, há uma diferença entre os termos clientes (ou beneficiários) e indivíduos. O indivíduo é único e identificado no plano através de um único código identificador. No entanto, cada indivíduo pode ter mais de um contrato, e por isso, terá duas carteirinhas que equivalem a dois códigos de beneficiários e será visto como cliente mais de

uma vez em um mesmo plano. Neste trabalho, nosso foco está nos clientes porque esses trazem consigo a informação do gasto gerado pelo indivíduo.

Uma técnica potencial para identificação dos perfis dos usuários (ou clientes) através da divisão dos mesmos em grupos é a análise de cluster descrita neste Capítulo 2. Porém, quando aplicadas com o intuito de separar os clientes de um plano de saúde, as técnicas usuais de análise de cluster não produzem resultados satisfatórios, uma vez que usuários com características similiares são alocados em diferentes grupos e usuários discrepantes entre si são alocados em um mesmo grupo. A alternativa para a qual recorremos é a decomposição em wavelets para realizar a análise de agrupamentos.

Vejamos porque isso acontece, utilizando como exemplo os clientes citados na Seção 2.1, na TAB.1: a figura abaixo ilustra a série temporal dos seus custos mensais.

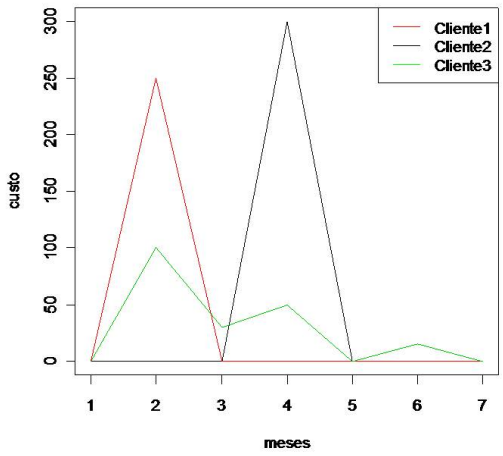


Figura 6: Série temporal do custo mensal (em reais) de três clientes

A distância euclidiana entre os clientes 1 e 2 é $390,52$ e é maior que a distância entre os clientes 1 e 3, essa última igual a $161,63$. No entanto, pela análise visual da série temporal, o cliente 1 é mais parecido com o cliente 2: ambos têm custo zero em quase todos os meses e apenas em certo mês eles têm algum custo. Ambos também obtiveram picos de tamanho relativamente próximos: 250 reis e 300 reais para os clientes 1 e 2, respectivamente. Portanto, os clientes 1 e 2 apresentam o mesmo padrão de comportamento da série e por isso, deveriam estar no mesmo grupo.

Porém, como os picos se encontram em momentos do tempo distintos, a distância euclidiana não consegue captar a "verdadeira distância" entre os clientes, já que essa é calculada "mês a

mês". Consequentemente, ao procedermos à análise de cluster, os clientes 1 e 2 estariam em clusters distintos.

Contudo, esperaríamos que clientes com o mesmo padrão de comportamento da série histórica dos custos estivessem em um mesmo grupo, ao passo que, clientes com diferentes padrões estivessem em clusters distintos. Em consonância com essa idéia, os clientes 1 e 2 estariam no mesmo cluster, e o cliente 3 deveria estar em um outro cluster, separado dos clientes 1 e 2.

Vemos, portanto, que a análise de clusters não é simples neste caso e a distância calculada com base nos custos dos clientes não é suficiente para segregá-los. A fim de resolver este problema, buscamos uma forma de reescrever os custos através dos coeficientes de wavelets, e depois aplicamos o método k-means aos custos transformados. No próximo capítulo traremos as definições das wavelets e os algoritmos de decomposição das funções através dos coeficientes de wavelets.

3 Wavelets

3.1 Introdução às wavelets

Segundo Morettin (1999), wavelet (ou ondaleta) é uma função capaz de decompor e descrever um sinal (ou uma outra função) no domínio da frequência, de forma a podermos analisá-lo em diferentes escalas de frequência e de tempo. A decomposição de uma função com o uso de wavelets é conhecida como transformada de wavelet e tem suas variantes contínua e discreta. Em análise de sinais, o termo domínio da frequência designa a análise de funções matemáticas com respeito à frequência, a qual indica o número de ocorrências de um evento (ciclos, voltas, oscilações, etc) em um determinado intervalo de tempo.

Os algoritmos de wavelets processam dados em diferentes escalas ou resoluções e, independentemente da função de interesse ser uma imagem, uma curva ou uma superfície, as wavelets oferecem uma técnica elegante na representação dos níveis de detalhes presentes (Cupertino, 2002). Elas constituem uma ferramenta matemática para decompor funções hierarquicamente, permitindo que uma função seja descrita em termos de uma forma grosseira, mais outra forma que apresenta detalhes que vão desde os menos delicados, aos mais finos. O resultado na análise de wavelets é "ver a floresta e as árvores".

Um sinal original ou uma função podem ser representados em termos de uma expansão em wavelets e as operações com dados podem ser feitas através de seus coeficientes. Se pudermos escolher as wavelets que melhor se adaptam aos dados, ou truncarmos os coeficientes menores do que um valor previamente estabelecido, os dados serão esparsamente representados. Essa "codificação esparsa" faz das ondaletas uma excelente ferramenta no campo de compressão de dados.

A idéia é que precisa-se de dois parâmetros: um parâmetro a, caracteriza a frequência, o outro, b, indica a posição do sinal. Famílias de funções $\psi_{a,b}$ definidas por

$$\psi_{a,b}(x) = |a|^{-1/2}\psi(\tfrac{x-b}{a});\ \text{a, b} \in \Re,$$

geradas a partir das operações de dilatação e translação da mesma função $\psi_{1,0}(x)$ ("wavelet mãe"), tornaram-se uma ferramenta muito importante em várias áreas da matemática pura e aplicada. Estas famílias são chamadas de wavelets.

3.2 Revisão bibliográfica e aplicabilidade da análise de wavelets

Embora a primeira menção tenha acontecido em 1909, por A. Haar, as wavelets de Haar ficaram no anonimato por muitos anos e, por um período muito longo, continuaram a ser a única base ortonormal de wavelets conhecida. Nos anos 30, usando a base de wavelets de Haar, Paul Lévy investigou o movimento Browniano. Ele mostrou que as funções da base de Haar eram melhores do que as da base de Fourier para estudar os pequenos e complicados detalhes do movimento Browniano.

Em processamento de sinais, trabalhos em técnicas entendidas como intimamente ligadas às wavelets começaram em 1976, por três pesquisadores franceses (A. Croisier, D. Esteban e C. Galand), os quais introduziram um banco de filtros que pode ser ser usado para decompor, fazer sub-amostragem e reconstruir um sinal. Uma década mais tarde, F. Mintzer, M. Smith e T. Barnwell construíram filtros que foram, posteriormente, relacionados com as bases de wavelets ortogonais.

Só recentemente, em 1985, Stephane Mallat deu às wavelets um grande impulso através de seu trabalho em processamento digital de imagens e, inspirado nos resultados de Mallat, Y. Meyer, construiu a primeira wavelet suave. Ao contrário das wavelets de Haar, as criadas por Meyer são continuamente diferenciáveis; contudo, elas não têm suportes compactos. Poucos anos mais tarde, Ingrid Daubechies usou os trabalhos de Mallat para construir um conjunto de bases ortonormais de wavelets suaves, com suportes compactos. Os trabalhos de Daubechies são os alicerces das aplicações atuais.

Contudo, podemos dizer que as ondaletas são um produto da colaboração de várias áreas, desde a matemática e física puras, até engenharia e processamento de sinais. A unificação de todos os pensamentos tornou-se um fator primordial para sua subsequente popularidade, impulsionando assim novas pesquisas na área. Wavelets são úteis em várias aplicações, como por exemplo: análise de sinais sísmicos (terremotos), análises de pressão sanguínea, ritmo cardíaco e ECG, análise de DNA e proteínas, modelagem geométrica, reconhecimento e síntese de fala, música, ressonância magnética, radar, redução de ruído e compressão.

3.3 De Fourier até Wavelets

Uma função pode ser convertida do domínio do tempo para o domínio da frequência através da transformada de Fourier, que decompõe uma função na soma de um número de componentes

senoidais multiplicados por coeficientes. Fourier foi o primeiro a estudar sistematicamente tal transformação, nomeada em sua honra como transformada de Fourier.

Apesar da funcionalidade da transformada de Fourier, existem muitas falhas nessa técnica. Através dela podemos extrair apenas informações sobre o domínio da frequência, mas estas frequências predominantes no sinal estão presentes em todos os instantes de tempo. Enquanto isso, na análise com wavelets, podemos extrair tanto as informações da função no domínio da frequência, quanto no domínio do tempo: a resolução ou detalhamento da análise no domínio da frequência diminui enquanto a resolução do tempo aumenta, sendo impossível aumentar o detalhamento em um dos domínios sem diminuí-lo no outro. Usando um análise wavelet, é possível escolher a melhor combinação dos detalhamentos para um objetivo estabelecido.

As funções seno e cosseno usadas na análise de Fourier não são locais e, portanto, desempenham uma tarefa muito pobre na aproximação de sinais muito localizados. A análise de Fourier é altamente instável em relação à presença de ruído nas funções devido ao caráter global.

3.4 Características

Para ser considerada uma wavelet, uma função precisa atender as seguintes características:

1. A área total sob a curva da função é 0, ou seja, $\int_{-\infty}^{\infty} \psi(x)\mathrm{d}x = 0$
2. A energia da função é finita, ou seja, $\int_{-\infty}^{\infty} \mid \psi(x) \mid^2 \mathrm{d}x < \infty$

A primeira característica acima sugere que $\psi(x)$ tende a oscilar acima e abaixo do eixo x. E a segunda característica revela que sua energia localiza-se em uma certa região (energia finita) e isso é o que diferencia as wavelets da análise de Fourier, já que essa última utiliza as funções sen(x) e cos(x), que são periódicas e com energia infinita.

A transformada contínua de wavelet $W(x)$ decompõe uma função definida no domínio do tempo em outra função, definida no domínio do tempo e no domínio da frequência:

$$W_{a,b}(x) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{a}} \psi(\frac{x-b}{a})\mathrm{d}x$$

que é uma função do parâmetro de dilatação a e do parâmetro de translação b. Se definirmos $\psi_{a,b}(x)$ como:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}} \psi(\frac{x-b}{a})\mathrm{d}x$$

entã podemos reescrever a transformada como o produto interno das funções $f(x)$ e $\psi_{a,b}(x)$:

$$W_{a,b}(x) = \langle f(x), \psi_{a,b}(x) \rangle = \int_{-\infty}^{\infty} f(x)\psi_{a,b}(x)dx.$$

Famílias de funções $\psi_{a,b}$ constituem uma base ortonormal para $\mathbb{L}^2$. O termo $\frac{1}{\sqrt{|a|}}$ é um fator de normalização, o qual garante que a energia de $\psi_{a,b}(x)$ seja independente de a e de b. Isto é, para todo a e b, temos:

$$\int_{-\infty}^{\infty} |\psi_{a,b}(x)|^2 dx = \int_{-\infty}^{\infty} |\psi(x)|^2 dx$$

No caso das wavelets discretas, os parâmetros de dilatação a e de translação b tomam apenas valores discretos. Para o parâmetro a, o mais usado na literatura (Cupertino, 2002 e Magalhães, 2007) são potências inteiras de um parâmetro de dilatação fixo $a_0 > 1$, isto é, $a = a_0^m$ (diferentes valores de m correspondem a wavelets de diferentes larguras). O parâmetro b discretizado depende de m: wavelets estreitas (alta frequência) são transladadas de pequenas distâncias a fim de cobrir todo o domínio espacial, enquanto que wavelets mais largas (baixa frequência) devem ser transladadas de uma distância maior. Visto que a largura de $\psi(a_0^{-m}x)$ é proporcional a a_0^m, escolhemos discretizar b por $b = nb_0a_0^m$, onde $b_0 > 0$ é fixado e $n \in \mathbb{Z}$. Portanto, $\psi_{m,n}(x)$ $= a_0^{-m/2}\psi(a_0^{-m}x - nb_0)$. É comum encontrarmos apenas os casos em que $a_0 = 1/2$, $b_0 = 1$, $n=$ k e $m= j$.

A transformada de wavelet discreta (DWT), do inglês *discrete wavelet transforms*, fornece informações suficientes tanto para a análise quanto para a síntese do sinal original, com uma redução significativa no tempo de computação, além de ser mais fácil de implementar, quando comparada à transformada de wavelet contínua. Os conceitos básicos das transformadas discretas serão introduzidos nas próximas seções, juntamente com as suas propriedades e os algoritmos usados para calculá-las.

Definição 3.4.1. *Define-se como wavelet mãe, ou simplesmente wavelet, uma função $\psi(x) \in \mathbb{Z}$, tal que a família de funções $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ onde j e k são inteiros arbitrários, seja uma base ortonormal para $\mathbb{L}^2$.*

Da definição acima, se ψ é uma wavelet, então $\psi_{j,k}$ também o será para qualquer $j, k \in \mathbb{Z}$ fixo e dizemos que o conjunto $\{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$ constitui uma base ortonormal de wavelets.

Para valores grandes de j, o fator de dilatação é grande e consequentemente a função $\psi(j)$, torna-se bastante espalhada. O parâmetro de escala ou dilatação é semelhante à escala utilizada em mapas. Como no caso dos mapas, escalas elevadas correspondem a uma visão não-detalhada

global (do sinal), e as escalas baixas correspondem a uma visão detalhada. De modo semelhante, em termos da frequência, as baixas frequências (escalas elevadas) correspondem a uma informação global do sinal (que geralmente se estende por todo o sinal), enquanto que as altas frequências (baixas escalas) correspondem a uma informação detalhada de um padrão escondido no sinal (que geralmente dura um tempo relativamente curto).

Em aplicações práticas, as escalas baixas (altas frequências) não duram por todo o sinal e escalas altas (baixas freqüências) normalmente duram por todo sinal. A escala, como uma operação matemática, ou dilata ou comprime um sinal. Escalas maiores correspondem à dilatação dos sinais e pequenas escalas correspondem a sinais comprimidos.

3.5 Wavelet de Haar

A Transformada de Haar é uma transformada matemática discreta usada no processamento e análise de sinais, na compressão de dados e em outras aplicações de engenharia e ciência da computação. Ela foi proposta em 1909 pelo matemático húngaro Alfred Haar. A transformada de Haar é um caso particular de transformada discreta de wavelet, definida como:

$$\psi(x) = \begin{cases} 1 & \text{se} \quad 0 \leq x < 1/2 \\ -1 & \text{se} \quad 1/2 \leq x < 1 \\ 0 & \text{caso contrário} \end{cases} \tag{5}$$

A função Haar ψ definida acima é chamada wavelet mãe (do inglês *"mother wavelet"*). A wavelet mãe "dá à luz" a toda uma família de wavelets, denominadas wavelets filhas, por meio de duas operações: dilatações e translações. Na FIG.7 vemos ilustrada a wavelet de Haar.

Denotamos as wavelets filhas por $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. O parâmetro j denota a compressão da função em torno do eixo x, enquanto o parâmetro k denota o efeito do deslocamento da função em torno do eixo x. Algumas dessas funções dilatadas e transladadas são representadas na FIG.8.

O suporte de $\psi_{j,k}$ é $[2^{-j}k, 2^{-j}(k+1))$, para $j \neq j'$ e $k \neq k'$. Fixada a escala j e tomando $k \neq k'$, as wavelets de Haar serão ortogonais, pois seus suportes não são coincidentes. Para

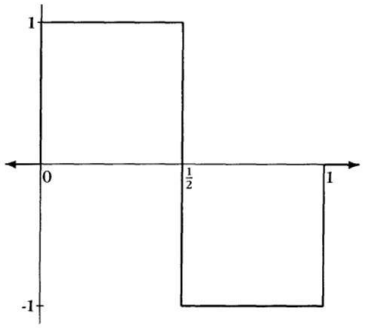


Figura 7: O gráfico de ψ dada pela equação(5)

escalas diferentes, é possível obter wavelets com suportes que se sobrepõem. É fácil mostrar que para $j < j'$, o suporte de $\psi_{j,k}$ está completamente dentro de uma região onde $\psi_{j',k}$ é constante. Neste caso, o produto interno entre as duas wavelets será proporcional à integral de ψ, que é zero. Desta forma mostra-se que wavelets em escalas diferentes são ortogonais, mesmo nos casos em que os suportes das funções se sobrepõem.

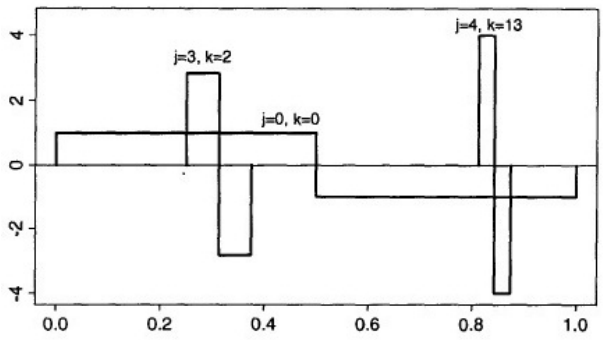


Figura 8: Exemplos de wavelets de Haar

A importância das wavelets é que quase toda função de importância prática pode ser bem aproximada por uma representação usando as wavelets. Isto é apresentado de maneira mais formal no Lema 3.5.1 a seguir (Cupertino, 2002) :

Lema 3.5.1. *Toda função $f \in \mathbb{L}^2$ pode ser arbitrariamente aproximada por uma combinação linear finita de $\psi_{j,k}$.*

Seja ϕ a função escala do intervalo [0, 1), associada à wavelet de Haar, isto é,

$$\phi(x) = \begin{cases} 1 & \text{se} \quad 0 \leq x < 1 \\ 0 & \qquad c.c \end{cases}$$

Como $\phi(2^j x - k)$ vale 1 no intervalo $[2^{-j}k, 2^{-j}(k+1))$ e zero, caso contrário, então $\{\phi_{j,k}(x)\}_{k\in\mathbb{Z}}$, com $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$, forma uma base ortonormal para V_j, o subespaço de $\mathbb{L}^2$, formado pelas funções constantes por partes em intervalos da forma $[2^{-j}k, 2^{-j}(k+1)]$, ou seja,

$$V_j = \{f : f(x) = \textstyle\sum_k a_{j,k}\phi_{j,k}(x), \sum_k \mid a_{j,k} \mid^2 < \infty \}.$$

Temos a seguinte inclusão: $V_j \subset V_{j+1}$, $\forall j \in \mathbb{Z}$. Dada arbitrariamente uma função $f \in \mathrm{V}_j$, ou seja, $f^j(x) = \sum_k a_{j,k}\phi_{j,k}(x)$, podemos escrever

$$f^j(x) \equiv f^{j-1} + \gamma^{j-1} \text{ onde } \gamma^{j-1} = \textstyle\sum_k d_{j-1,k}\psi_{j-1,k}(x)$$

onde $f^{j-1} \in V_{j-1}$, $\gamma^{j-1} \in \mathrm{W}_{j-1}$ e W_j é subespaço de $\mathbb{L}^2$ gerado por $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$, ou seja,

$$W_j = \{f \in \mathbb{L}^2 : f(x) = \textstyle\sum_k d_{j,k}\psi_{j,k}(x), \sum_k \mid d_{j,k} \mid^2 < \infty \}.$$

A decomposição da função dada acima pode ser vista da seguinte maneira: ao passarmos de V_{j-1} para V_j obtemos uma versão de maior resolução da função (aumentamos sua resolução por um fator de 2) e os detalhes (incrementos de informações), são representados por wavelets de W_j. Em outras palavras, à medida que aumentamos o parâmetro j, a amplitude do intervalo das funções $\phi(x)$ e $\psi(x)$ diminui pela metade, de forma que quanto maior o valor de j, melhor é aproximação. A FIG.9 contém a função original e suas aproximações para $j = 4$, 5 e 6. Vemos que para j igual a 6, temos uma melhor aproximação da função.

As informações passíveis de serem representadas por wavelets em uma dada escala constituem uma faixa, de forma que podemos imaginar escalas subseqüentes como faixas adjacentes, compreendendo detalhes de diferentes tamanhos. Dessa forma, dizemos que as wavelets são ferramentas que guardam a informação de detalhes correspondentes a passagens de duas resoluções ou escalas consecutivas. Os espaços $W_j's$ são espaços de detalhes.

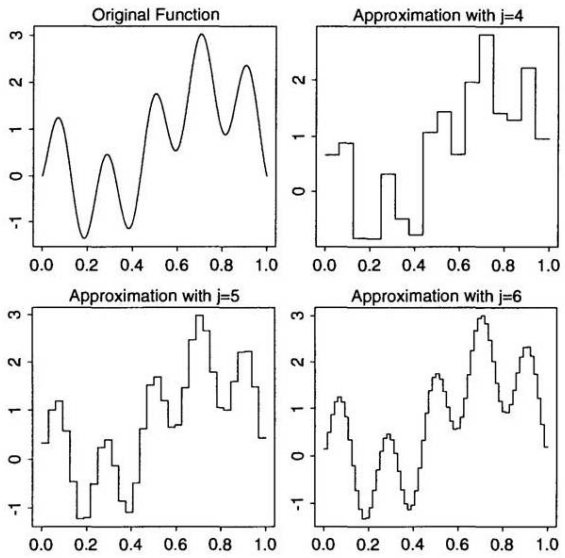


Figura 9: Função original e sua aproximação a cada resolução
Fonte: OGDEN, R. T. Essential wavelets for statistical applications and data analysis. Department of Statistics ,University of South Carolina, Columbia, p.13–28, 1965.

Funções de wavelets e funções escala são ortogonais: $\langle \psi_{j,k}, \phi_{j',k'} \rangle = 0$, para todo j', k', j, k inteiros, logo, V_j e W_j são mutuamente ortogonais para todo j. Como $W_{j-1} \subset \mathrm{V}_j$, segue-se que W_{j-1} é complemento ortogonal de V_{j-1} em relação a V_j.

Podemos escrever uma função $f^j \subset \mathrm{V}_j$ como a soma de sua alta resolução, $f^{j+J} \subset \mathrm{V}_{j+J}$ cuja resolução é 2^J vezes maior do que a versão original, f^j, mais detalhes $\gamma^{j+l} \subset \mathrm{W}_{j+l}$, $l = 1...J$, correspondentes às escalas intermediárias, os quais são representados por wavelets.

A wavelet de Haar é a única wavelet com suporte compacto, para a qual se tem uma forma analítica fechada para os coeficientes. Existem também as wavelets de Daubechies que têm suportes compactos e podem ser tomadas tão suaves e com quantos momentos nulos quanto desejamos; entretanto, não se conhece uma forma analítica fechada para os coeficientes e por isso, são calculados numericamente. As wavelets de Haar fornecem um paradigma para todas as demais wavelets e é importante manter em mente que tudo o que for desenvolvido neste trabalho têm aplicação muito mais ampla: todos os princípios a serem discutidos referentes às wavelets de Haar geralmente se aplicam para as demais wavelets ortogonais.

A transformada de Haar pode ser usada para representar um grande número de funções $f(x)$ como sendo o somatório:

$$f(x) = \sum_{k\in\mathbb{Z}} a_k \phi(x-k) + \sum_{j=0}^{\infty} \sum_{k\in\mathbb{Z}} d_{j,k} \psi(2^j x - k)$$

sendo que a_k e $d_{j,k}$ são os parâmetros a serem calculados.

3.6 Cálculo rápido dos coeficientes de wavelets na base de Haar

Para a wavelet de Haar temos as seguintes relações:

$$\psi(x) = \sqrt{2}\left(\frac{1}{\sqrt{2}}\phi(2x) - \frac{1}{\sqrt{2}}\phi(2x-1)\right) \equiv \sqrt{2}(h_0\phi(2x) + h_1\phi(2x-1)) \tag{6}$$

$$\phi(x) = \sqrt{2}(\frac{1}{\sqrt{2}}\phi(2x) + \frac{1}{\sqrt{2}}\phi(2x-1)) \equiv \sqrt{2}(g_0\psi(2x) + g_1\psi(2x-1)) \tag{7}$$

Podemos generalizar as Equações (6) e (7):

$$\phi(x) = \sqrt{2}\left(\sum_k h_k\phi(2x-k)\right) \tag{8}$$

$$\psi(x) = \sqrt{2}\left(\sum_k g_k\phi(2x-k)\right) \tag{9}$$

Podemos expressar $\phi_{j-1,k}(x)$ em termos de $\phi_{j,k}(x)$

$$\phi_{j,k}(x) = 2^{j-1/2}\phi(2^{(j-1)}x - k) = \textstyle\sum_n h_n\phi_{j,n+2k}$$

Os coeficientes h_k's e g_k's são chamados de coeficientes dos filtros da função de escala e da wavelet, respectivamente. Mostra-se que esses coeficientes determinam completamente as funções, ou seja, tudo o que é necessário para uma análise wavelet são os coeficientes dos filtros. Apenas conhecê-los é suficiente para determinar o valor da função em qualquer ponto, com a precisão desejada, através de um algoritmo recursivo. Sabe-se ainda que os coeficientes g_k podem ser determinados a partir de h_k e vice-versa. Além disso, eles não podem assumir quaisquer valores. Uma primeira restrição pode ser obtida, bastando para tanto integrar ambos os lados das Equações (8) e (9). Lembrando ainda que busca-se uma base ortonormal e, portanto, a norma $\mathbb{L}^2$ das funções de base é unitária. Temos que:

$$\int_{-\infty}^{\infty} |\psi(x)|^2 \mathrm{d}x = \int_{-\infty}^{\infty} |\sqrt{2}\sum_k h_k\phi(2x-k)|^2 \mathrm{d}x \tag{10}$$

$$1 = \sum_k h_k^2 \int_{-\infty}^{\infty} |\sqrt{2}\phi(2x-k)|^2 \mathrm{d}x, \tag{11}$$

sendo $\int_{-\infty}^{\infty} |\sqrt{2}\phi(2x-k)|^2 \mathrm{d}t = 1$

$$1 = \sum_k h_k^2 \tag{12}$$

Da mesma forma, para a Equação (9) obtém-se:

$$1 = \sum_k g_k^2 \tag{13}$$

Definimos o coeficiente de wavelet como $d_{j,k}$ e a "média"$a_{j,k}$ como as projeções de f sobre $\psi_{j,k}$ e $\phi_{j,k}$, respectivamente, ou seja, $a_{j,k}$= $\langle f, \phi_{j,k} \rangle$, e $d_{j,k}$= $\langle f, \psi_{j,k} \rangle$:

$$a_{j-1,k} = \frac{1}{\sqrt{2}}(a_{j,2k} + a_{j,2k+1}), \tag{14}$$

$$d_{j-1,k} = \frac{1}{\sqrt{2}}(a_{j,2k} - a_{j,2k+1}), \tag{15}$$

as quais nos permitem o cálculo rápido dos coeficientes de wavelets $d_{j-1,k}$ a partir dos coeficientes $a_{j-1,k}$, onde j_0 pode ser visto como uma escala grosseira, tal que a projeção de f sobre o espaço das funções constantes em intervalos da forma $[2^{-j_0}k, 2^{-j_0}(k+1))$, ou seja, $\sum_k a_{j_0,k}\phi_{j_0,k}(x)$ seja uma boa aproximação para f.

Esses coeficientes dependem somente do comportamento local de $f(x)$ no intervalo descrito acima. Esta é uma diferença das séries de Fourier ou integrais de Fourier, nas quais cada coeficiente depende do comportamento global de f. O coeficiente $a_{j,k}$ captura a média de f e $d_{j,k}$ captura as mudanças em f.

Veremos a seguir que, para uma wavelet em geral, temos as seguintes relações:

$$a_{j-1,k} = \sum_n h_n a_{j,n+2k} \tag{16}$$

$$d_{j-1,k} = \sum_n g_n a_{j,n+2k} \tag{17}$$

As relações (16) e (17) nos dão um algoritmo rápido de decomposição de uma função. So-

mando e subtraindo as Equações (14) e (15), temos as seguintes fórmulas que fornecem um algoritmo de reconstrução da função,

$$a_{j,2k} = \frac{1}{\sqrt{2}}(a_{j-1,k} + d_{j-1,k}) \tag{18}$$

$$a_{j,2k+1} = \frac{1}{\sqrt{2}}(a_{j-1,k} - d_{j-1,k}) \tag{19}$$

Os coeficientes de escala são fornecidos pelo produto interno de f com as correspondentes funções de base.

$$a_{j,k} = \langle f, \phi_{j,k} \rangle = \int f(x)\phi_{j,k}dx = \int_{k2^{-j}}^{(k+1)2^{-j}} f(x)2^{j/2}dx \tag{20}$$

Em termos novamente dos coeficientes de Haar, temos por exemplo:

$$a_{0,2} = \int_2^3 f(x)\mathrm{d}x = \sqrt{2}(\int_2^{2,5} f(x)\sqrt{2}\mathrm{d}x + \int_{2,5}^3 f(x)\sqrt{2}\mathrm{d}x) = (a_{1,4} + a_{1,5})/\sqrt{2} \tag{21}$$

Podemos generalizar e derivar os coeficientes de escala:

$$a_{j,k} = (a_{j+1,2k} + a_{j+1,2k+1})/\sqrt{2} \tag{22}$$

Dessa forma, a expressão (22) pode ser usada recursivamente para computar todos os coeficientes de escala, do nível mais alto para o nível mais baixo. Os coeficientes de wavelets também podem ser generalizados da seguinte maneira:

$$d_{j,k} = (a_{j+1,2k} - a_{j+1,2k+1})/\sqrt{2} \tag{23}$$

Podemos construir os coeficientes de filtros da função wavelet em termos dos coeficientes de filtros da função escala mais próxima:

$$g_n = (-1)^n h_{1-n}.$$

Como vimos, uma função de base na escala j pode ser obtida a partir de funções de base na escala anterior $j+1$. Também é possível obter a representação de uma função na escala $j+1$ utilizando, para tanto, a representação em j.

O processo mostrado acima pode ser visto como uma forma de se obter representações cada vez mais finas $(f_1, f_2, f_3, ...)$ da função original f, sendo as informações (detalhes) armazenadas

em forma de wavelets $(\gamma_1, \gamma_2, \gamma_3, ...)$.Também construiu-se aqui, de forma intuitiva, uma análise em resoluções múltiplas, em que cada passo da decomposição é, na verdade, uma projeção feita sobre um subespaço de menor resolução, de forma que a seqüência de subespaços formada será uma seqüência encaixante.

Na prática, dada uma coleção arbitrária de $n = 2^J$ valores, que representa o total da amostra a qual devemos considerar, temos que j varia de 0 a $J-1$, e k varia de 0 a $2^j - 1$. Nos problemas de interesse, a representação em termos de coeficientes de wavelets é esparsa no sentido de que a maioria dos coeficientes $d_{j,k}$ são nulos ou muito pequenos e por isso, podemos ignorá-los - daí a idéia de compressão por trás da representação em bases de wavelets.

3.7 Análise de resolução em escalas múltiplas

Neste capítulo daremos a definição da análise de resolução em escalas múltiplas, que abreviaremos por ARM, a qual foi formulada por Meyer em 1986. Ela fornece um referencial onde bases de wavelets são naturalmente compreendidas, bem como permite a construção de novas bases. Através dela podemos ver as bases ortonormais de wavelets como uma ferramenta para descrever matematicamente o "incremento na informação" necessário para se ir de uma aproximação grosseira (com menor resolução) para uma aproximação mais fina (com maior resolução). Um sinal pode ser visto como uma componente suave acrescido de flutuações (detalhes). A distinção entre o que é suave e o que são detalhes é feita de acordo com o nível de resolução empregado. Uma análise em resoluções múltiplas (ARM) é uma forma de se representar uma função em diferentes resoluções.

Na Seção 3.5, introduzimos uma idéia de ARM para o caso particular das wavelets de Haar.

Observação 3.7.1. *Na literatura, a terminologia "escala", "nível", e ocasionalmente "resolução" são, algumas vezes, usadas de forma intercambiável. Nesta dissertação, o termo nível de resolução expressa a quantidade de informação envolvida na análise de multiresolução e também corresponde ao subespaço que contém a função f . O termo escala será usado para designar a quantidade 2^{-j}. Além disso, j largo corresponde à uma fina escala (ou pequena escala), enquanto j pequeno corresponde à uma escala mais grosseira (ou grande escala). Quanto maior for o parâmetro j, maior é o nível de resolução, mais fina é a aproximação e temos maior riqueza de detalhes. Quanto menor for o parâmetro j, menor é o nível de resolução, mais grosseira é a aproximação e temos menos detalhes.*

Seguem as definições e o teorema acerca da análise de resolução em escalas múltiplas, dados por Cupertino (2002):

Definição 3.7.1. *Uma ARM é uma sequência, $\{V_j\}_{j\in\mathbb{Z}}$, de subespaços de $\mathbb{R}$, representando os sucessivos níveis de resoluções, tal que satisfaça as seguintes condições:*

1. $\ldots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \ldots$
2. $f \in V_j$ *se, e somente se,* $f(2^j.) \in V_{j+1}$
3. $f \in V_0$ *implica que* $f(. - k) \in V_0$ *para todo* $k \in \mathbb{Z}$, *e* $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ *forma uma uma base ortonormal para* V_0.
4. $\bigcap_{j\in Z} V_j = \{0\}$
5. $\overline{\bigcup_{j\in Z} V_j} = \mathbb{L}^2$

A sequência de espaços $(V_j)_{j\in\mathbb{Z}}$ representa uma sequência de subespaços encaixantes. Cada subespaço V_j consiste de funções que são constantes por partes em intervalos exatamente duas vezes menores que V_{j-1}. Todas as construções de wavelets, com exceção de alguns casos patológicos, têm como ponto de partida a estrutura acima, chamada de análise de resolução em escalas múltiplas. A figura abaixo é representativa dos espaços encaixantes.

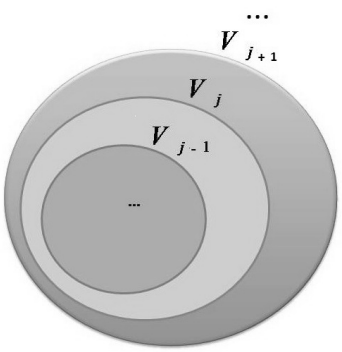


Figura 10: Espaços Encaixantes

Teorema 3.7.1. *Se uma sequência de subespaços $(V_j)_{j\in\mathbb{Z}}$ e f satisfazem às condições acima, então existe uma base ortonormal de wavelets $\{\psi_{j,k}|j,k \in Z\}$ para $\mathbb{L}^2$, tal que:*

$$P_{j+1} = P_j + \sum_k \langle f, \psi_{j,k}\rangle \psi_{j,k} \tag{24}$$

onde P_j é a projeção ortogonal sobre V_j.

O item 2 da Definição 3.7.1 expressa que todos os espaços estão relacionados por escala a um mesmo espaço V_0, e por isso, cada aproximação pode ser escrita como uma soma de uma aproximação mais grosseira e os detalhes. Ainda por causa desta propriedade, se $f(x) \in V_j$, então $f(x-k) \in V_j$, para todo $k \in \mathbb{Z}$. As condições 2 e 3 implicam que $\{\phi_{j,k}\}_{j,k\in\mathbb{Z}}$ é uma base ortonormal para V_j para todo $j \in \mathbb{Z}$. A condição 5 assegura:

$$\lim_{n\to-\infty} P_j f = f = \sum_k \langle f, \phi_{j,k}\rangle \phi_{j,k} \tag{25}$$

para todo $f \in \mathbb{L}^2$.

Como vimos, qualquer função $f \in \mathbb{L}^2$ pode ser aproximaçada por uma função constante por partes f^j, e quanto maior o nível de resolução j, melhor a aproximação. A figura 12 ilustra a função suavizada e suas três aproximações. A cada nível j, uma função f^j é construída como uma aproximação da função original, a qual pode ser descrita como uma soma da aproximação grosseira mais próxima, f^{j-1} e da função de detalhes γ^{j-1}. Cada detalhe pode ser escrito como uma combinação linear de wavelets $\psi_{j,k}$.

Uma importante propriedade da multiresolução ou ARM pode ser escrita como:

$$V_j = V_{j-1} \oplus W_{j-1} \tag{26}$$

onde A $\oplus$ B nos diz que um subespaço A é complemento ortogonal do outro subespaço B. Adicionalmente, W_j é um subespaço criado pelas wavelets, e a mesma propriedade **??** da Definição 3.7.1 válida para o espaço V_j também é válida para W_j:

$$f \in W_j \text{ se, e somente se, } f(2^j.) \in W_{j+1}$$

A Equação (26), conjuntamente ao Teorema 3.7.1 expressam a principal filosofia da análise de wavelets: é possível construir uma aproximação a cada nível de resolução como uma combinação linear das dilatações e translações da função escala ϕ, e as diferenças entre as aproximações são expressadas como uma combinação linear das dilatações e translações da função wavelet ψ. Como já foi dito, as funções escala e wavelet são ortogonais. Os subespaços W_j e $W_{j'}$ são ortogonais, para $j \neq j'$.

A figura acima fornece uma demonstração dessa aproximação, na qual cada uma delas pode ser escrita como uma combinação linear das funções de base $\phi_{j,k}$.

$$
\begin{array}{ccccccccc}
\cdots & & W_{j-1} & & W_j & & W_{j+1} & & \cdots \\
& \searrow & & \searrow & & \searrow & & \searrow & \\
\cdots & \rightarrow & V_{j-1} & \rightarrow & V_j & \rightarrow & V_{j+1} & \rightarrow & \cdots
\end{array}
$$

Figura 11: Relação dos espaços de aproximação e espaços de detalhes

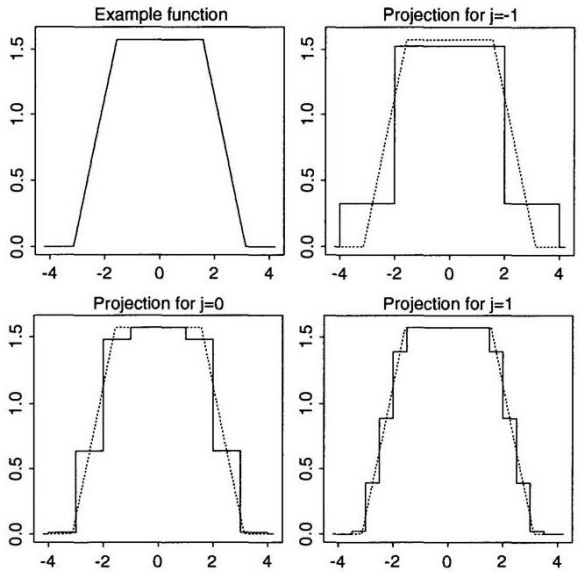


Figura 12: Um exemplo de uma função e suas aproximações em diferentes níveis
Fonte: OGDEN, R. T. Essential wavelets for statistical applications and data analysis. Department of Statistics ,University of South Carolina, Columbia, p.13–28, 1965.

Cada coeficiente do nível j é visto como uma diferença entre os coeficientes do nível $j+1$. Contudo, o principal objetivo da análise de multiresolução é escrever o sinal em termos de componentes. Busca-se uma parcimoniosa representação que preserva as características da função original, mas que expressa a função em termos de um pequeno conjunto de coeficientes.

A cada passo em que o nível de resolução cresce, movemos de uma aproximação grosseira e uma aproximação mais fina é criada. A análise consiste em estudar os detalhes presentes no sinal (ou função), ou diferenças na aproximação feita em cada nível de resolução adjacente.

3.8 Algoritmos rápidos de decomposição e reconstrução de uma função

Na Seção 3.6, havíamos descrito algoritmos rápidos para se calcular os coeficientes de wavelets de uma função, para o caso particular da wavelet de Haar. Neste capítulo obteremos algoritmos

rápidos para se fazer a decomposição (análise) e reconstrução de uma função.

Conforme mencionado, exceto para as wavelets de Haar, todas as famílias ortonormais de wavelets de suportes compactos, como por exemplo, as wavelets de Daubechies, symmlet, coiflet, entre outras, e suas funções escalas não possuem uma forma analítica fechada. Nestes casos, seus valores têm que ser calculados numericamente. Transcrevendo as relações descritas em 22 e 23:

$$\phi(x) = \sqrt{2}\left(\sum_k h_k \phi(2x-k) \right) \tag{27}$$

$$\psi(x) = \sqrt{2}\left(\sum_k g_k \phi(2x-k) \right) \tag{28}$$

Note que, de (27), temos:

$$\phi_{j-1,k}(x) = 2^{j-1/2}\phi(2^{j-1}x-k) = 2^{(j-1)/2}\left(\sum_n h_n \phi(2^{(j-1)/2}x-2k-n) \right) = \sum_n h_n \phi_{j,2k+n}(x), \forall j \in n. \tag{29}$$

De maneira análoga,

$$\psi_{j-1,k}(x) = \sum_n g_n \phi_{j,2k+n}(x) \tag{30}$$

Como P_j é a projeção ortogonal sobre V_j, temos que

$$P_j f = \sum_k a_{j,k}\phi_{j,k} \tag{31}$$

A análise de wavelets agora procede na direção de j decrescente. Descreveremos o passo $j \to j-1$: assuma que os coeficientes $\{a_{j,k}\}_k$ sejam conhecidos e estejam armazenados numa matriz. De (27) e (29), temos:

$$a_{j-1,k} = \left\langle P_{j-1}f, \phi_{j-1,k} \right\rangle = \langle f, \phi_{j-1,k} \rangle = \langle f, \sum_n h_n \phi_{j,2k+n} \rangle = \sum_n h_n \langle f, \phi_{j,2k+n} \rangle = \sum_n h_n a_{j,2k+n} \tag{32}$$

Seja Q_j a projeção ortogonal sobre W_j, então, $\langle f, \psi_{j,k} \rangle = \langle Q_j f, \psi_{j,k} \rangle$ assim, definindo $d_{j,k}$ $= \langle f, \psi_{j,k} \rangle$, temos:

$$Q_j f = \sum_k d_{j,k}\psi_{j,k} \tag{33}$$

E, além disso:

$$d_{j-1,k} = \langle Q_{j-1}f, \psi_{j-1,n}\rangle = \langle f, \psi_{j-1,k}\rangle = \langle f, \sum_n g_n \phi_{j,2k+n}\rangle = \sum_n g_n \langle f, \phi_{j,2k+n}\rangle = \sum_k g_n a_{j,2k+n}, \tag{34}$$

o que nos dá a seguinte recursão:

$$d_{j-1,k} = \sum_n g_{n-2k} a_{j,n} \tag{35}$$

As fórmulas construídas acima são uma forma de se obter os coeficientes que representam a função numa escala mais grosseira, a partir de uma versão de alta resolução e os detalhes. Na passagem $j \longrightarrow j-1$, perde-se a resolução por um fator de 2. A nova versão de baixa resolução de f , que é a projeção de f sobre V_{j-1}, é obtida a partir dos coeficientes $a_{j-1,n}$ e os detalhes correspondentes a esta perda, ou seja, a diferença das projeções de f sobre V_{j-1} e V_j, respectivamente, são armazenadas nos coeficientes $d_{j-1,n}$. A aplicação que leva uma função f nos seus coeficientes de wavelets $d_{j-1,k}$ é geralmente referida como transformada discreta de wavelets, calculada a partir do algoritmo de decomposição.

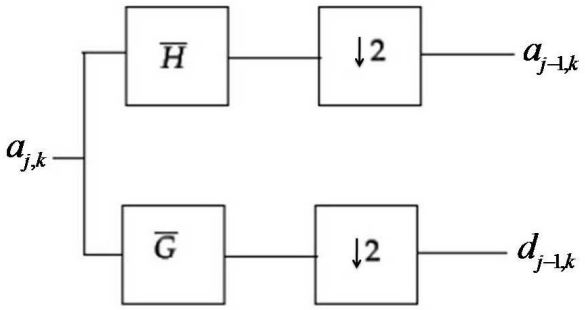


Figura 13: Esquema representando um passo da transformada de wavelet rápida (decomposição ou análise) em termos dos filtros

Para tanto, é necessário apenas conhecermos os coeficientes $a_{j,k}$, que representam a função f numa dada escala j e também os coeficientes de filtro h_n da função escala associada à análise.

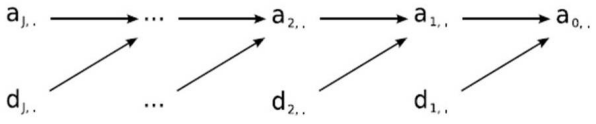


Figura 14: Algoritmo rápido de decomposição

A cada passo que a resolução diminiu, o número de coeficientes reduz pela metade. Tomando como exemplo o esquema ilustrado na figura abaixo, o nível de resolução 3 é o mais alto e com o maior número de coeficientes.

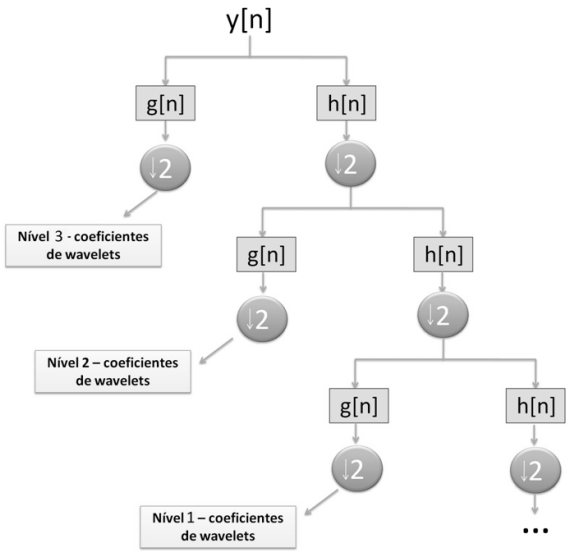


Figura 15: Esquema representando os filtros de escala e de wavelets

A figura a seguir fornece uma síntese do algoritmo conhecido por reconstrução, pois deseja-se ser capaz de reconstruir a função original, partindo de uma baixa resolução para a alta resolução. Percebe-se que esse caminho traçado pelo algoritmo de reconstrução é o inverso do percorrido pelo algoritmo de decomposição. Constrói-se desta maneira, um algoritmo rápido para passar de uma escala para outra subsequente, a qual terá mais coeficientes e por conseguinte, mais detalhes.

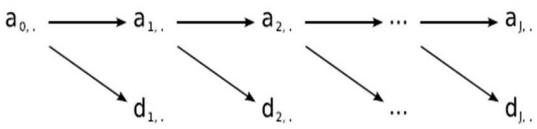


Figura 16: Algoritmo rápido de reconstrução

Tomando a projeção do sinal f sobre o subespaço V_{j+1}, onde $V_{j+1} = V_j \oplus \mathrm{W}_j$, pode-se escrever a projeção da seguinte forma:

$$P_{V_{j+1}} f = P_{V_j} f + P_{W_j} f \tag{36}$$

$$\sum_n a_{j+1,n}\phi_{j+1,n} = \sum_k a_{j,k}\phi_{j,k} + \sum_k d_{j,k}\psi_{j,k} \tag{37}$$

Podemos aplicar as wavelets à uma sequência ou vetor de dados: $y = (y_1, y_2, ...y_n)$, onde cada y_i é um número real, sendo i variando de 1 a n. Para obter os coeficientes, assumimos que o tamanho da sequência n é múltiplo de dois. Ou seja, $n = 2^J$. O maior nível de resolução é consiste em $n/2 = 2^{J-1}$ observações. O menor nível de resolução será 0, que equivale a um coeficiente. Dessa forma, teremos sempre no mínimo um coeficiente e, no máximo, o equivalente à metade do tamanho da amostra.

A escolha do nível é subjetiva e depende do interesse do pesquisador. Mas sabemos que, quanto maior o nível de resolução j, maior o número de coeficientes e melhor a aproximação. A seguir temos uma tabela com um exemplo geral, onde y= $(1, 1, 7, 9, 2, 8, 8, 6)$. 2^J=8, e portanto, j varia de 0 a $J - 1$.

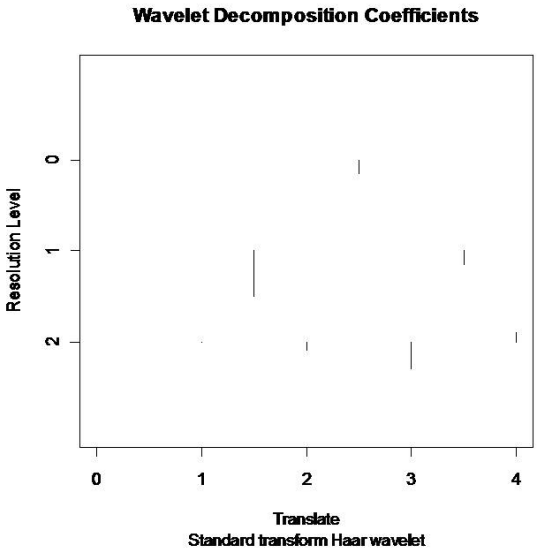
Tabela 2: Coeficientes

Nível 2	$d_{2,0}$	$d_{2,1}$	$d_{2,2}$	$d_{2,3}$
	0.000000	-1,414214	-4,242641	1,414214
Nível 1	$d_{1,0}$	$d_{1,1}$		
	-7	-2		
Nível 0	$d_{0,0}$			
	-2,12132			

Explicando de uma forma mais grosseira, para um mesmo nível de resolução, cada coeficiente calculado corresponde a um determinado intervalo da função original. Este intervalo depende também do parâmetro k, sendo que à medida que aumentamos o valor de k, os intervalos são deslocados percorrendo todo o domínio da função.

O gráfico a seguir fornece um plot dos coeficientes de wavelets. Os coeficientes $d_{j,k}$ são plotados da escala mais fina até a escala mais grosseira (topo do gráfico). Os valores dos coeficientes são exibidos por uma marca vertical localizada ao longo de uma linha central imaginária presente em cada nível. Assim, as três marcas localizadas no nível 2 correspondem aos três coeficientes $d_{2,1}$, $d_{2,2}$, $d_{2,3}$. O coeficiente $d_{2,0}$ não é plotado, por ser igual a zero. O parâmetro de localização k é rotulado "Translate" e indica a posição aproximada na sequência original a partir da qual os coeficientes são derivados.

Neste trabalho atentaremos somente para os coeficientes de wavelets $d_{j,k}$, pois eles captam



os detalhes ou as mudanças nas funções, o que vai de encontro ao nosso interesse com relação às séries de custos. Esperamos que os coeficientes de wavelets identifiquem as mudanças nessas séries, isto é, os meses com custos mais baixos e os meses com picos ou custos mais elevados.

4 Análise de cluster associada às wavelets aplicada aos dados de uma operadora

Os dados utilizados neste trabalho são de uma determinada operadora de planos de saúde, a qual possui 99.865 mil clientes cadastrados e identificados com um código de beneficiário. Este número é único e preserva algumas características dos indivíduos, tais como data de nascimento, sexo, data de início do contrato, data do fim do contrato (caso o cliente tenha saído do plano) e titularidade (se o cliente é titular ou dependente). Temos também o custo mensal de cada cliente, de agosto de 2003 a novembro de 2008, em um total de 64 meses, os quais também podemos ver como características. Esses dados estão dispostos em uma matriz, de forma que cada cliente se encontra em uma linha, e em cada coluna temos suas características.

Nosso objetivo aqui proposto é identificar quantos tipos de clientes existem na carteira do plano de saúde, baseando-se nas séries históricas dos custos. Sendo assim, atentaremos apenas para os custos, e a matriz com os dados será então composta por um total 99.865 mil linhas (que correspondem a 99.865 clientes) e 64 colunas com os custos mensais dos respectivos clientes.

Conforme explicamos na Seção 2.5, os métodos de análise de cluster, quando aplicados à carteira em questão, não conseguem separá-la em grupos homogêneos, uma vez que clientes com mesmos padrões das séries de custos são alocados em diferentes grupos. Por outro lado, clientes com diferentes padrões são alocados em um mesmo grupo. Isso ocorre porque a distância entre os elementos da amostra é calculada "ponto a ponto"ou "mês a mês". E, além disso, as séries de custos apresentam comportamento bastante oscilatório e não estacionário, o que dificulta a identificação dos perfis.

Por isso, buscamos um método para reescrevê-las, a qual possibilite a aplicação da análise de cluster e produza grupos com características semelhantes. Encontramos nas wavelets uma possível solução para o problema, visto que elas são uma forma de reescrever qualquer função e permitem a análise de fenômenos oscilatórios, não estacionários e variantes no tempo, características essas semelhantes às identificadas nas séries de custos.

A idéia geral é que não precisamos de todos os coeficientes de wavelets para caracterizar uma função, mas apenas de alguns que resumem toda a informação contida nela. Cada coeficiente é calculado com base em um intervalo da função original que dependerá do nível de resolução adotado. Quanto maior o nível de resolução, menor o intervalo da função original considerado para o cálculo de cada coeficiente.

Em suma, podemos enxergar cada série de custo como uma função e reescrevê-la em termos dos coeficientes de wavelets. A nova matriz com dados conterá cada cliente em uma linha e em cada coluna, teremos seus coeficientes. Ressalta-se que perderemos o momento exato em que ocorreram os custos e até mesmo, os picos (meses com custos mais elevados em relação ao padrão da série), o que condiz com nosso interesse, uma vez que basta apenas termos uma idéia da localização temporal dos picos.

Em regiões de suavidade da função, os coeficientes serão pequenos e em regiões com picos, os coeficientes serão elevados. Esperamos que o método de análise de cluster via wavelets forneça uma melhor separação dos grupos e consigamos identificar os perfis de clientes.

Denota-se o custo em cada mês como $cust_t$, para t variando de 1 a 64 e o vetor de custos de cada cliente i como:

$$Cliente_i = (Cust_{i1}, Cust_{i2}, ..., Cust_{it}), \text{ para } t = (1, 2, ...64) \text{ e } i = (1, 2, ...99865).$$

Observamos problemas na frequência de utilização do plano nos primeiros nove meses de custos, por isso, criamos pseudo-dados da seguinte forma: replicamos os nove últimos meses de custos nos primeiros nove meses. Ou seja, os custos mensais de cada cliente para $t = (1, 2....9)$ foram substituídos pelos custos mensais para $t = (56, 57...64)$. Então, os custos $(Cust_{i1}, Cust_{i2}, ..., Cust_{i9})$ são iguais a $(Cust_{i56}, Cust_{i57}, ..., Cust_{i64})$.

Ao invés de calcularmos os coeficientes de wavelets com base nos custos, optamos por calculá-los em termos do logaritmo dos custos. Procedemos assim em todas as análises descritas na Seção 4.1, referente aos procedimentos metodológicos. A apresentação dos dados nessa escala é útil pelo fato de termos uma gama de valores de custos e o logaritmo reduz a representação a uma escala mais fácil de ser visualizada e manejada. Nos meses em que os custos eram iguais a zero, somamos o valor de uma unidade para que o logaritmo resultante fosse zero. Para o melhor entendimento sobre o porquê utilizamos a escala logarítmica, temos o seguinte exemplo:

Um cliente A teve um custo no mês t igual a 1.000 reais e no mês $t + 1$ seu custo foi igual a 3.000 reais; enquanto o cliente B teve um custo de 30.000 reais no mês t e de 28.000 reais no mês $t + 1$. A diferença de custo de um mês para o outro para ambos os clientes é de 2.000 reais. Embora a diferença seja a mesma, para o cliente A, o custo triplicou de um mês para o outro, e por isso, o aumento de 2.000 reais foi mais significativo para esse cliente do que para o cliente B.

Quando calculamos a diferença entre os logaritmos dos custos referentes aos meses t e $t + 1$, o resultado é 1,098 para o cliente A e 0,0689 para o cliente B. Dessa forma, a escala logarítmica

expressa a diferença relativa entre os custos mensais dos clientes A e B e revela o impacto do aumento do custo de um mês para o outro para cada um deles, sendo esse impacto maior para o cliente A. Veja a FIG.17 abaixo:

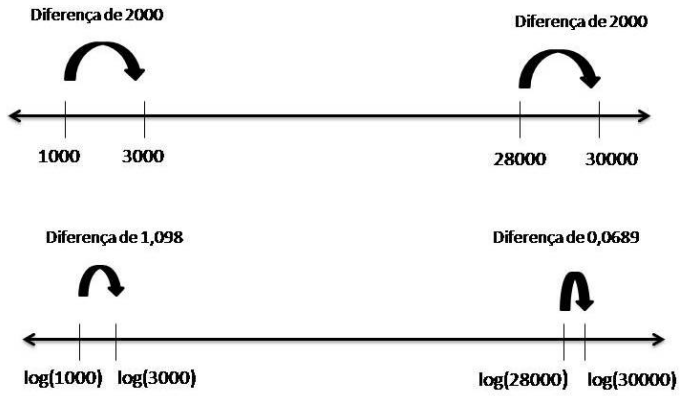


Figura 17: Esquema com as diferenças dos custos mensais e do logaritmo dos custos mensais

Nota-se que, quanto maiores os custos nos meses t e $t+1$, menor será a diferença entre seus respectivos logaritmos. E, quanto menores os custos nos meses t e $t+1$, maior será a diferença entre os logaritmos.

Denotamos o vetor com os logaritmos dos custos de cada cliente i como:

$$Cliente_i = (log(Cust_{i1}), log(Cust_{i2}), ..., log(Cust_{it})),\ t = (1, 2, ...64) \text{ e } i = (1, 2, ...99865).$$

A seguir temos a TAB.3, que contém a frequência de clientes que apresentaram algum custo em cada mês, ou seja, $cust_t > 0\ \forall\ t$. Quando o cliente realiza algum procedimento médico, seja consulta, exame, terapia, internação, entre outros, ele gera custos para a operadora. Cada procedimento possui um custo específico que depende de sua complexidade, e o custo mensal (ou anual) do cliente dependerá dos procedimentos realizados e da frequência de utilização.

A base de dados em questão não possui clientes que tiveram custo zero nos 64 meses, ou seja, $cust_t = 0\ \forall\ t$. Esses clientes já representam um perfil a ser avaliado pela operadora. Mas, lembramos que um cliente pode obter custo zero em um ou mais meses, mas terá custo maior que zero em pelo menos um mês.

Vemos que, em cada ano, os meses de dezembro, janeiro e fevereiro apresentam menor número de clientes com $cust_t > 0$, o que já é observado em diversas operadoras de planos de saúde por serem meses tipicamente de férias e em geral, a demanda por serviços de saúde diminui nesses meses. Junho, julho e agosto são meses com maior número de clientes com $cust_t > 0$, devido ao

Tabela 3: Frequência de clientes que apresentaram algum custo de ago/2003 a nov/2008

ago/03	set/03	out/03	nov/03	dez/03							
22104	22728	21395	22049	23639							
jan/04	fev/04	mar/04	abr/04	mai/04	jun/04	jul/04	ago/04	set/04	out/04	nov/04	dez/04
21757	21811	22376	20563	29693	35461	35842	37102	35515	34510	35972	32805
jan/05	fev/05	mar/05	abr/05	mai/05	jun/05	jul/05	ago/05	set/05	out/05	nov/05	dez/05
32496	29535	33348	31032	30707	30448	28626	31813	29042	27886	27791	25314
jan/06	fev/06	mar/06	abr/06	mai/06	jun/06	jul/06	ago/06	set/06	out/06	nov/06	dez/06
26841	25475	30090	26524	28970	27641	28923	28771	26190	26400	25719	22679
jan/07	fev/07	mar/07	abr/07	mai/07	jun/07	jul/07	ago/07	set/07	out/07	nov/07	dez/07
24297	23768	27012	24393	25295	24432	24654	25590	22658	24998	22951	19381
jan/08	fev/08	mar/08	abr/08	mai/08	jun/08	jul/08	ago/08	set/08	out/08	nov/08	
21608	21239	22104	22728	21395	22049	23639	21757	21811	22376	20563	

inverno que ocorre nesse período e traz consigo o aumento de doenças respiratórias, tais como: gripe, resfriado, pneumonia, asma, bronquite, entre outras.

A TAB.4 fornece o número de coeficientes que teremos em cada nível de resolução. Podemos enxergar cada linha da matriz dos dados como um vetor de tamanho 64 referentes aos 64 meses de custos. Por isso, teremos no máximo seis níveis de resolução, sendo que o último nível terá 32 coeficientes de wavelets que correspondem à metade do número de observações do vetor de custos de cada cliente.

Tabela 4: Resolução e o número de coeficientes

Resolução	Total de coeficientes
Nível 5	32
Nível 4	16
Nível 3	8
Nível 2	4
Nível 1	2
Nível 0	1

As ondaletas foram implementadas através do pacote *"wavethresh"* disponível no *software* R. Denotamos o vetor de coeficientes de cada cliente i referente ao nível 5 de resolução como:

$$Cliente_i = (Coef_{i1}, Coef_{i2}, ..., Coef_{it}),\ t = (1, 2, ...32) \text{ e } i = (1, 2, ...99865).$$

Após reescrevermos os custos em termos dos coeficientes, realizamos a análise de cluster através do mesmo *software*, uma vez que o método *K-means* já está implementado nele. Conforme explicamos na Seção 2.4, *a priori*, o método k-means requer a especificação do número

de grupos. Por esta razão, implementamos tal método considerando de um a quinze grupos e, posteriormente, calcularemos a medida de homogeneidade definida como critério de escolha do número de grupos final.

Em cada grupo ou perfil resultante, buscamos observar padrões de comportamento das séries temporais dos custos dos clientes segundo algumas características: tamanho e quantidade de picos, tempo de permanência em custo maior que zero, tempo de permanência em custo igual a zero, presença ou ausência de picos consecutivos, custo total, entre outras. Tais passos foram seguidos em todas análises descritas na próxima seção, apenas alteramos os vetores com as características dos clientes.

4.1 Procedimentos metodológicos

Em uma primeira análise, reescrevemos os logaritmos dos custos mensais de cada cliente em termos dos 32 coeficientes de wavelets correspondentes ao nível de resolução máximo. Posteriormente, realizamos a análise de cluster na qual as distâncias entre os clientes foram calculadas em relação a cada um desses coeficientes.

Essa análise proporcionou uma melhor separação dos grupos em comparação às análises baseadas somente nos custos dos clientes. Todavia, em cada grupo, encontramos clientes com características não similares entre si e com custo total discrepante do custo total médio do grupo. Clientes com custo total elevado (acima de 20 mil reais) não foram segregados dos clientes com custo total baixo (abaixo de 500 reais). Por esses motivos, concluímos que as wavelets, por si só, ainda não foram suficientes para produzir os perfis de clientes.

Por conseguinte, realizamos uma segunda análise a qual associamos os 32 coeficientes de wavelets ao logaritmo do custo total dos clientes. Assim, para cada cliente i temos o seguinte vetor de observações:

$$Cliente_i = (Coef_{i1}, Coef_{i2}...Coef_{it}, Log(Custototal_i)),\ t = (1, 2, ...32) \text{ e } i = (1, 2, ...99865).$$

onde $Custototal_i$ é igual a $Cust_{i1} + Cust_{i2} + ... + Cust_{it}$, para $t = (1, 2, ...64)$, e representa a soma dos custos mensais de cada cliente i em 64 meses, e o $Log(Custototal_i)$ representa o logaritmo dessa soma.

Também realizamos uma terceira análise na qual consideramos os coeficientes de wavelets (32 coeficientes no total) associados ao logaritmo do custo total, esse último disposto repetidamente

em 32 colunas. Procedendo dessa forma, damos pesos iguais a $0,5$ para os coeficientes e $0,5$ para o logaritmo do custo total. Para cada cliente temos um vetor de observações com um total de 64 variáveis:

$$Cliente_i = (Coef_{i1}, Coef_{i2}...Coef_{it},$$
$$\text{Log(Custo total}_{i1}), Log(Custototal_{i2}), ...Log(Custototal_{ij})),\ t = (1, 2, ...32) \text{ e } j = (1, 2, ...32).$$

sendo que $Log(Custototal_{i1}) = Log(Custototal_{i2})\ ... = Log(Custototal_{ij}))$.

Nenhuma das análises descritas apresentaram resultados favoráveis, uma vez que não conseguimos identificar grupos compostos por clientes com séries temporais de custos similares. Além disso, a distribuição do custo total em cada grupo se mostrou muito discrepante e concentrada.

Posteriormente, incorporamos duas novas variáveis aos coeficientes de wavelets e ao logaritmo do custo total disposto repetidamente em 16 colunas : número de meses nos quais o cliente obtém custo maior que zero (a qual denominaremos cont) e número máximo de meses consecutivos nos quais o cliente permanece em custo igual a zero (a qual denominaremos seq). Portanto, temos um total de 50 colunas ou variáveis para cada vetor de observações dos clientes, sendo que para cada uma delas os pesos dados foram: $0,64$ para as wavelets; $0,32$ para o logaritmo do custo total e $0,02$ para cada uma das variáveis restantes. Priorizamos dar maior peso para as wavelets em comparação às demais variáveis, embora também demos um peso considerável para o logaritmo do custo total. O vetor de observações de cada cliente i é dado por:

$$Cliente_i = (Coef_{i1}, Coef_{i2}...Coef_{it}, Log(Custototal_{i1}),$$
$$\text{Log(Custo total}_{i2}), ...Log(Custototal_{ij}), cont_i, seq_i),\ t = (1, 2, ...32),\ j = (1, 2, ...16) \text{ e } i =$$
$$(1, 2, ...99865).$$

Nessa análise conseguimos identificar os perfis de clientes, e em cada grupo formado, observamos poucos clientes com características não similares entre si e em proporção muito menor em relação às análises realizadas anteriormente. A distribuição do custo total em cada grupo ainda apresentou assimetria. No entanto, ainda não ficamos satisfeitos com os resultados, e buscamos reescrever os custos de outra forma, a qual explicaremos a seguir.

5 Discussão dos resultados

Para definir a forma de reescrever os custos, partimos do pressuposto de que não necessitamos do momento exato em que ocorreram os picos presentes nas séries históricas dos custos, mas sim do tamanho e da quantidade de picos. Optamos por exemplificar as séries a seguir com base nos custos, e não no logaritmo dos custos, apenas para facilitar a visualização dos gráficos.

Suponhamos os exemplos na FIG.18 que correspondem a um único perfil de cliente. Temos duas séries de custo em um total de 64 meses, mas com apenas um pico de 100 reais em cada uma delas, e nos 63 meses restantes, os custos são iguais a zero. Porém, os picos encontram-se em momentos distintos do tempo: na série 1, o pico localiza-se no mês 3, e na série 2, o pico localiza-se no mês 41. Contudo, os dois clientes apresentam o mesmo comportamento da série de custos: custo total baixo, maior tempo de permanência em custo zero, apenas um pico com custo mais elevado.

Figura 18: Exemplos de clientes com mesmo padrão de comportamento da série de custos

Se reescrevermos os custos do cliente 1 (representado na série 1) e os custos do cliente 2 (representado na série 2), em termos dos coeficientes de wavelets correspondentes a cada nível de resolução, teremos os mesmos valores independente do nível escolhido. O que mudará é a ordem dos coeficientes, pois ela depende de onde se encontram os picos e do comportamento geral das séries. Também sabemos que os coeficientes de wavelets podem ser negativos, o que a nosso ver não interfere nos resultados. Por isso, optamos por considerar somente os valores absolutos dos coeficientes e, em seguida, ordená-los de forma decrescente.

Para ambas as séries da FIG.18, teremos apenas um coeficiente no nível 0, dois coeficientes no nível 1, até o nível 5, o qual teremos 32 coeficientes e corresponde ao nível máximo.

No entanto, podemos considerar todos os níveis de resolução possíveis, uma vez que cada um deles resume a informação contida na função (ou na série de custos) com maior ou menor riqueza de detalhes. Quanto menor o nível de resolução (nível zero), mais grosseira é a aproximação e menos detalhes temos da função original. Quanto maior o nível de resolução (nível 5), mais detalhes temos da função original e mais fina é a aproximação.

Explicando de uma forma mais grosseira, o nível zero corresponde a uma visão mais geral das séries de custos, no qual teremos apenas um coeficiente e seu valor equivale, aproximadamente, a uma média global dos custos em todos os meses. O nível cinco, também de uma forma grosseira, corresponde a uma média dos pares de custos. Por exemplo, o primeiro coeficiente desse nível é aproximadamente uma média dos custos nos dois primeiros meses de análise, o segundo coeficiente é aproximadamente uma média dos custos no terceiro e no quarto meses de análise, e assim sucessivamente.

Os valores dos coeficientes da série 1 em cada nível de resolução são:

Nivel 0: 12,5

Nível 1: 17,67767 0

Nível 2: 25 0 0 0

Nível 3: 35,35534 0 0 0 0 0 0 0

Nível 4: -50 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Nível 5: 0 70,71068 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

E, os coeficientes da série 2 em cada nível de resolução são:

Nível 0: - 12,5

Nível 1: 0 17,67767

Nível 2: 0 0 -25 0

Nível 3: 0 0 0 0 0 35,35534 0 0

Nível 4: 0 0 0 0 0 0 0 0 0 0 50 0 0 0 0 0

Nível 5: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 70,71068 0 0 0 0 0 0 0 0 0 0

Observe que os valores absolutos dos coeficientes em cada nível são iguais e o que muda é a ordem em que eles aparecem nas séries. Tomando o valor absoluto e ordenando os coeficientes do maior para o menor, teremos vetores de coeficientes iguais para ambas as séries:

Nível 0: 12,5

Nível 1: 17,67767 0

Nível 2: 25 0 0 0

Nível 3: 35,35534 0 0 0 0 0 0 0

Nível 4: 50 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Nível 5: 70,71068 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Desse modo, o vetor de coeficientes para cada cliente i será dado pela junção dos coeficientes de todos os níveis, começando do nível zero ao nível máximo:

$$Cliente_i = (Coef_{i1}, Coef_{i2}, ..., Coef_{it}),\ t = (1, 2, ...63) = \text{e}\ i = (1, 2)$$

$Cliente_i$=$(12,5; 17,67767; 0; 25; 0; 0; 0; 35,35,534; 0; 0; 0; 0; 0; 0; 0; 50; 0; 0;$

0;0;0;0;0;0;0;0;0;0;0;0;0;70,71068;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0)

Ao calcularmos a distância euclidiana entre os clientes 1 e 2 com base nos vetores formados pelos coeficientes, essa será zero. E, assim, ambos estarão em um mesmo grupo na análise de cluster. Clientes cuja distância entre eles é pequena são alocados para um mesmo grupo, e clientes cuja distância entre eles é maior, são alocados para grupos diferentes.

Na FIG.19, também temos mais dois exemplos de clientes com mesmo perfil do comportamento da série de custos: ambos permanecem em custo igual a 100 reais por 30 meses consecutivos, e nos 34 meses restantes, os custos são iguais a zero. No entanto, o primeiro cliente permanece nos 30 primeiros meses consecutivos em custo igual a 100 reais, enquanto o segundo cliente permanece nos 30 últimos meses consecutivos em custo igual a 100 reais.

Figura 19: Exemplos de clientes com mesmo padrão de comportamento da série de custos

Ao calcularmos os valores absolutos dos coeficientes de cada nível e ordená-los de forma decrescente, eles serão iguais para os dois clientes e a distância entre os mesmos também será zero. O mesmo ocorrerá se calcularmos esses coeficientes em termos dos logaritmos dos custos mensais.

Nível 0: 375

Nivel 1: 35,35534 0

Nivel 2: 50 0 0 0

Nivel 3: 70,71068 0 0 0 0 0 0 0

Nivel 4: 100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Nivel 5: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

O vetor formado pelos coeficientes de cada um desses clientes será dado por:

$$Cliente_i = (Coef_{i1}, Coef_{i2}, ..., Coef_{it}),\ t = (1, 2, ...63) \text{ e } i = (1, 2)$$

$$Cliente_i = (375; 35,35534; 0; 50; 0; 0; 0; 70,71068; 0; 0; 0; 0; 0; 0; 0; 100; 0; 0; 0; 0; 0; 0;$$

$$0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0)$$

Além dos coeficientes de wavelets, também consideramos o logaritmo do custo total de cada cliente, uma vez que ele tem um peso importante na definição dos perfis. Realizamos diversas análises com o logaritmo dos custos totais disposto nas colunas e quando os dispomos repetidamente em dezesseis colunas, os clusters resultantes nos fornecem características bem delineadas das séries de custos. Ressalta-se que subtraimos uma constante dos logaritmos dos custos totais afim de que estejam na mesma escala das wavelets: os logaritmos encontram-se entre 1,16 e 14,15, e os coeficientes de wavelets encontram-se entre 0 e 23,23.

Em suma, primeiramente reescrevemos as séries dos logaritmos dos custos de cada cliente da mesma forma que executamos nos exemplos dados acima. Obtivemos um total de 63 coeficientes de wavelets para cada cliente, começando do nível mais baixo para o nível mais alto, e os associamos ao logaritmo do custo total disposto repetidamente no vetor de observações de cada cliente i. Dessa forma, demos peso 0,20 para o logaritmo dos custos totais e 0,80 para os coeficientes de wavelets. A base de dados final têm, portanto, um total de 99.865 linhas correspondentes aos clientes, e para cada um deles, temos 79 colunas ou variáveis. As distâncias euclidianas entre os clientes foram calculadas com base nessas 79 variáveis. A explicação sobre porque utilizamos o logaritmo dos custos se encontra no capítulo 4.

O vetor com as observações de cada cliente i é dado por:

$$Cliente_i = (Coef_{i1}, Coef_{i2}, ..., Coef_{it}, Log(Custototal_{i1}), ...Log(Custototal_{ij})),$$

$$t = (1, 2, ...63),\ i = (1, 2, ...99865) \text{ e } j = (1, 2, ...16).$$

Finalmente, prosseguimos à análise de cluster através do método *k-means* e construímos a medida de homogeneidade (FIG.20), usada como critério de escolha do número final de grupos. Não consideramos seu valor no gráfico quando temos apenas um cluster, devido à variabilidade ser extremamente elevada quanto temos apenas um cluster.

À medida que o número de clusters aumenta, a homogeneidade reduz devido ao aumento da variabilidade entre os grupos. Mas, a partir de cinco clusters, não observamos um decaimento acentuado da homogeneidade. Por isso, optamos em permanecer com este número de grupos, nos quais cada um deles representa um perfil de cliente.

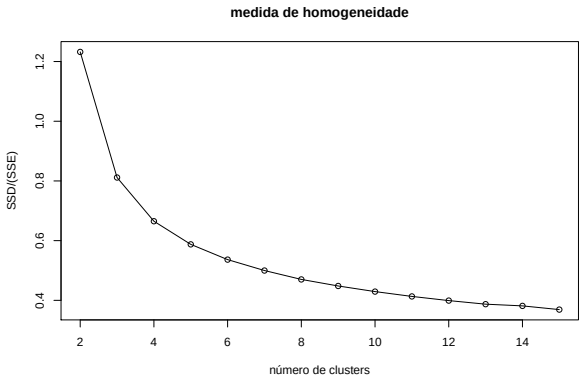


Figura 20: Medida de homogeneidade

A numeração dos clusters descritos a seguir foi obtida de acordo com o *software* R. Cada cliente possui uma identificação referente ao cluster para o qual foi alocado, sendo este número único. Apresentaremos algumas descritivas da carteira do plano de saúde, para posteriormente, caracterizarmos os perfis de clientes resultantes da análise de cluster.

Conforme já citamos, temos um total de 99.685 clientes, dos quais cerca de 60% são mulheres e 40% são homens. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE, 1998), estima-se que um quarto da população do país é coberta por pelo menos um plano de saúde, sendo essa porcentagem maior para as mulheres. Podemos, portanto, ter uma possível explicação para a maior proporção de mulheres na carteira em questão.

A idade média dos clientes é 35,56 anos, e temos clientes com idade mínima de zero anos e com idade máxima de 97 anos. Quanto à distribuição da idade por sexo (FIG.21), a idade mediana das mulheres é superior à idade mediana dos homens. Metade das mulheres tem idade inferior a 36 anos, e metade dos homens tem idade inferior a 32 anos. Portanto, nota-se que a carteira não é envelhecida, uma vez que a idade mediana é 35 anos e apenas 25% dos clientes têm idade superior a 53 anos.

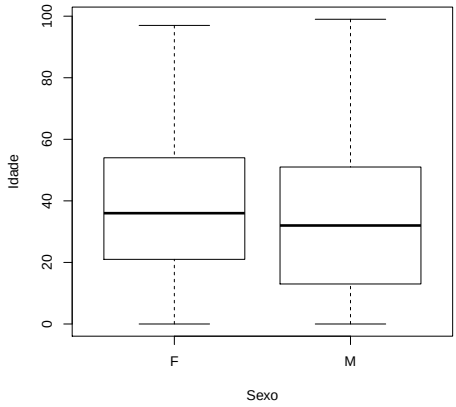


Figura 21: Distribuição da carteira por idade e por sexo

Quanto à titularidade, 42% dos clientes são titulares, 30% são filhos (ou filhas) e 19% são cônjuges, e os 9% restantes correspondem aos demais tipos de titularidade, tais como pai, mãe, agregado, entre outros. No que tange ao custo, o custo total mínimo é R$3,20 e o custo total máximo é R$ 1.399.000,00. O custo per capita é R$5.599,00 e metade dos clientes possui custo total inferior a R$2.147,00. Em síntese, vemos que a distribuição do custo total é bastante assimétrica, concentrada e discrepante. Segue o sumário do custo total da carteira:

Tabela 5: Sumário do custo total

Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
R$3,20	R$682,40	R$2.147,00	R$5.599,00	R$5.228,00	R$1.399.000,00

Temos também a distribuição do custo total por sexo: as mulheres apresentam maior custo médio e maior mediana. O custo total mediano para as mulheres é 2.625,00 reais e 1.578,00 para os homens. As mulheres da carteira são mais velhas em comparação aos homens, e geram maiores gastos para a operadora. Segue a descrição dos perfis de clientes e em cada um deles apresentaremos exemplos das séries históricas dos custos de alguns clientes. E, nos gráficos com as séries de custos, o eixo das abscissas corresponde aos custos mensais em reais.

Cluster I

O cluster I é composto por 21.848 clientes, que representam 21% do total de clientes. A proporção de clientes é 56% para o sexo feminino e 44% para o sexo masculino. 39% dos clientes

são titulares e 37% são filhos ou filhas, e os demais são pais, ou mães, entre outros tipos de titularidade. Os clientes têm, em média, 29,35 anos e custo per capita igual a R$ 1.061,00 segundo menor custo per capita em comparação aos demais grupos descritos. O custo total mínimo é 244,30 reais e o custo total máximo é 20.410,00 reais, e a mediana é 728,10 reais. O custo per capita das mulheres é 1.114,00 reais, enquanto o custo per capita dos homens é 992,30 reais.

Tabela 6: Sumário do custo total - Cluster I

Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
R$244,30	R$505,80	R$728,10	R$1.061,00	R$1.152,00	R$20.410,00

Apenas a idade média do cluster V é inferior à idade média do grupo I. Vemos que o grupo com menor custo per capita (grupo V) é aquele com menor idade média e o grupo com segundo menor custo per capita (grupo I) é aquele com a segunda menor idade média. A idade mediana das mulheres (29 anos) é superior à idade mediana dos homens (27 anos), e a distribuição das idades para ambos os sexos é simétrica. Portanto, observamos que quanto menor o custo per capita do grupo, menor é a idade média.

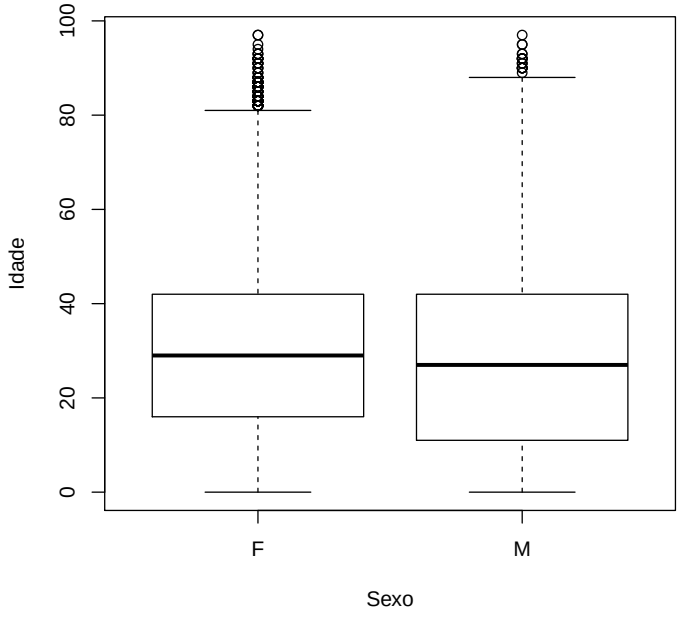


Figura 22: Distribuição etária por sexo - Cluster I

As séries dos custos possuem até seis picos cujos tamanhos variam entre 80 e 200 reais. Em geral, tais picos correspondem ao total de picos de toda a série. Apenas o cluster V possui menor

quantidade de picos e de menor tamanho em relação ao cluster I, embora o tamanho dos picos seja próximo para ambos os clusters.

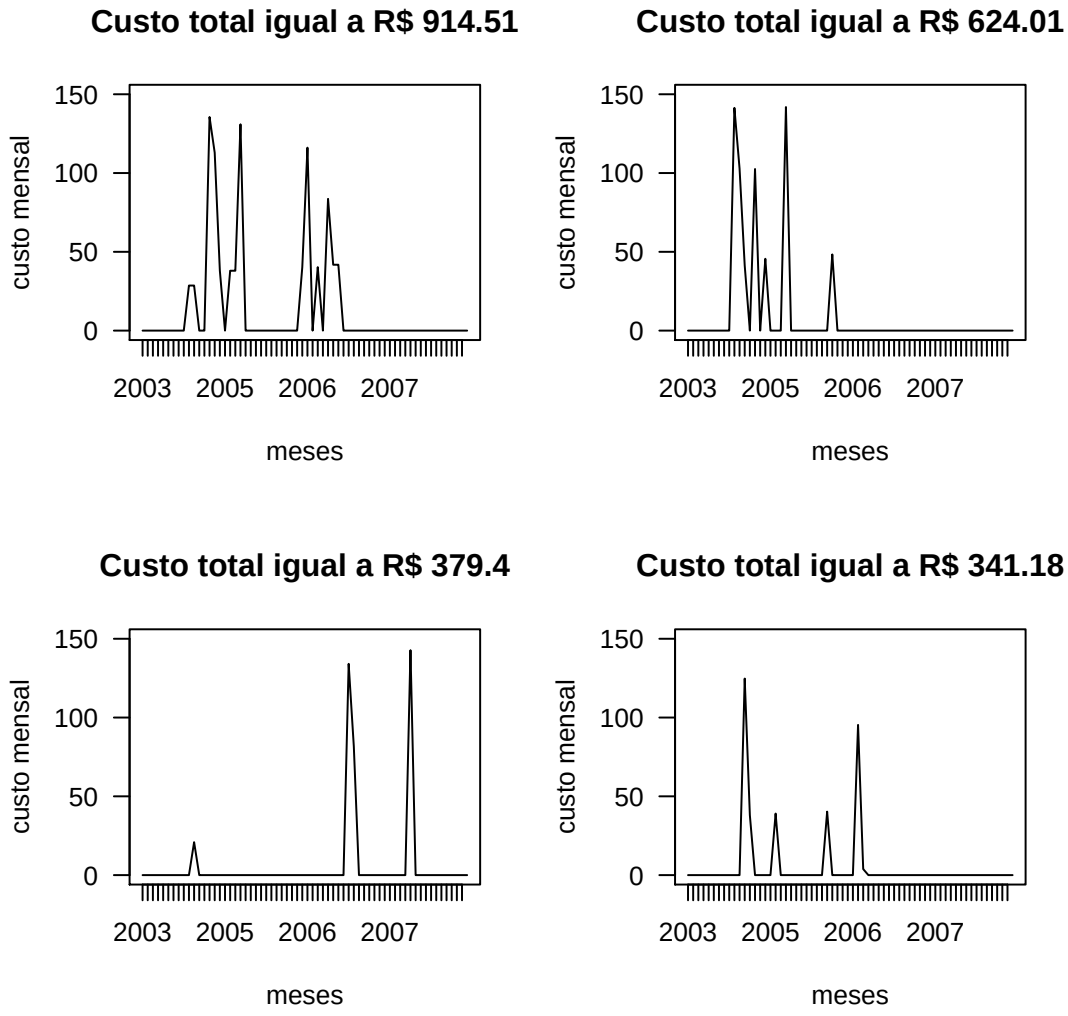


Figura 23: Exemplos de séries de custos - Cluster I

Observa-se também um maior tempo de permanência em custo zero em relação aos demais grupos, exceto quando também comparamos com grupo V.

Temos, portanto, um grupo com clientes jovens, que realizam procedimentos médicos eventualmente, cujos custos totais são baixos e que resultam em menor quantidade e tamanho de picos presentes nas séries históricas.

Cluster II

O cluster II é composto por 10.482 clientes, os quais representam apenas 10% do total de clientes que compõem o plano de saúde. Observamos uma maior porcentagem de mulheres (cerca de 64%) em relação à porcentagem de homens e no que diz respeito à titularidade, metade dos clientes do grupo são titulares.

Enquanto o grupo V é composto pelos clientes com os menores custos da carteira, o grupo II é composto pelos clientes mais caros, ou que geram maiores despesas assistenciais para a operadora. O custo per capita é 28.840,00 reais, o custo total mínimo e o custo total máximo são iguais a R$ 4.032,00 e R$1.399.000,00, respectivamente. Metade dos clientes têm custo total inferior à R$15.520,00. O cliente mais caro (com maior custo total) do plano também foi alocado para o cluster II.

Tabela 7: Sumário do custo total - Cluster II

Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
R$4.032,00	R$9.360,00	R$15.520,00	R$28.840,00	R$28.870,00	R$1.399.000,00

A idade média do grupo é 53,02 anos. Vemos através da FIG.24 que a idade mediana dos homens é superior à idade mediana das mulheres. O custo per capita das mulheres é 25.480 reais e o custo per capita dos homens é 34.710 reais. Neste grupo, notamos o oposto do que ocorreu nos demais grupos: os homens obtiveram maior custo mediano e maior idade mediana. O grupo II contém os clientes com idade mais avançada da carteira, sendo que os homens geram gastos maiores para o plano do que as mulheres.

As séries históricas dos custos possuem picos cujos tamanhos variam entre 1.000 e 10.000 reais. Também observamos séries com picos de 25.000 e 40.000 reais. Além disso, observa-se que os meses com custo elevado ocorrem consecutivamente por até três meses. Os tamanhos dos picos das séries deste grupo são os maiores em relação a todos os grupos. O tempo de permanência em custo maior que zero é de 16,91 meses, e o tempo de permanência em custo igual a zero é de 20,75 meses consecutivos.

O grupo II é caracterizado por um custo total elevado, os quais são gastos em apenas alguns meses. Sendo assim, torna-se complicada alguma ação por parte da operadora, uma vez que não sabemos quando ocorrerão esses gastos extremamente elevados. No entanto, também observamos clientes que gastam em média, 500 reais mensais, que somados para todos os meses, geram gastos

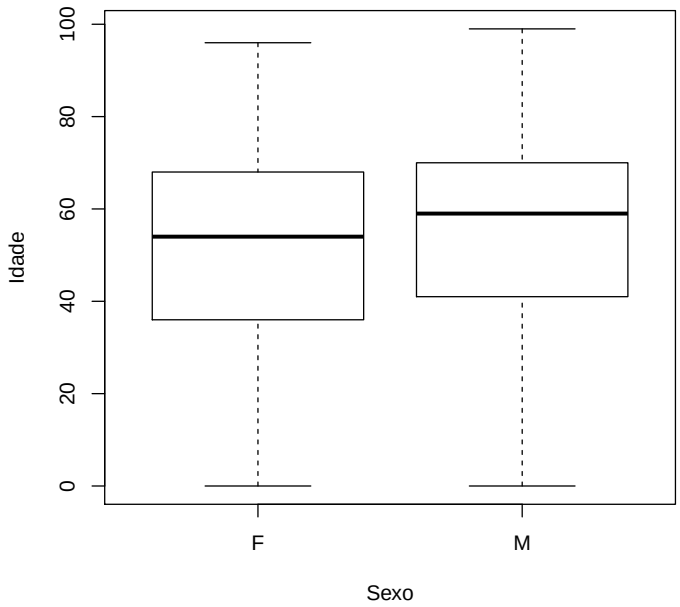


Figura 24: Distribuição etária por sexo - Cluster II

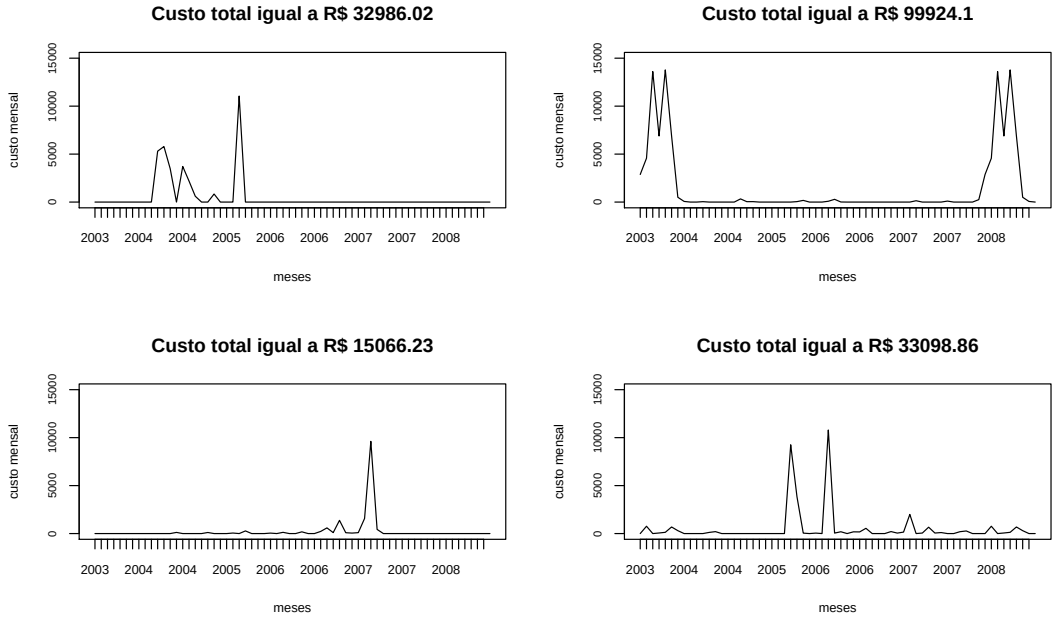


Figura 25: Exemplos de séries de custos - Cluster II

totais elevados. Para esses, a operadora consegue realizar intervenções com o intuito de reduzir seus custos e melhorar a saúde dos clientes.

É provável que os clientes desse grupo realizem procedimentos caros e mais complexos, tais como internações, cirurgias, os quais são, em média, mais caros que as consultas ambulatoriais e em consultórios médicos, exames, entre outros procedimentos de menor complexidade.

Cluster III

O cluster III é composto por 27.429 clientes, sendo que 71% desse total são mulheres e apenas 29% são homens. Esse é o maior grupo no que diz respeito à quantidade de clientes, os quais representam 27% do total de clientes da amostra, e também é o grupo com maior percentual de mulheres.

Quanto aos tipos de titularidade mais relevantes deste grupo, temos que aproximadamente metade dos clientes são titulares e 23% são cônjuges. A idade média do grupo é 43,4 anos, a idade média das mulheres e dos homens é 44,41 e 40,96, respectivamente. Vemos através da FIG.26 a maior variabilidade presente na distribuição etária dos homens.

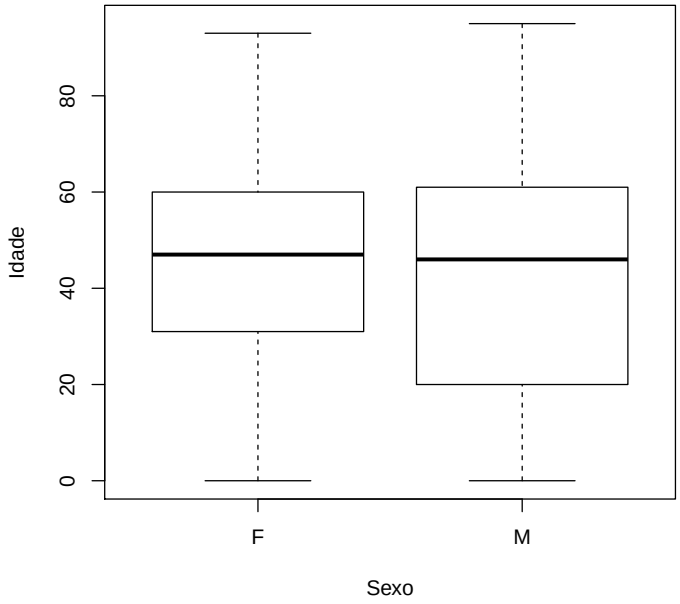


Figura 26: Distribuição etária por sexo - Cluster III

O custo total é, em média, 6.320 reais (TAB.8). O custo total mínimo e o custo total máximo são iguais a 1.277 reais e 49.020 reais, respectivamente. Metade dos clientes do grupo III possuem custo total superior à 5.202 reais.

Tabela 8: Sumário do custo total - Cluster III

Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
R$1.277,00	R$3.633,00	R$5.202,00	R$6.320,00	R$7.758,00	R$49.020,00

Vemos através do box-plot que a distribuição do custo total apresenta uma pequena assimetria para ambos os sexos e notamos a presença de outliers. As mulheres possuem custo total mediano um pouco maior em relação ao custo mediano dos homens.

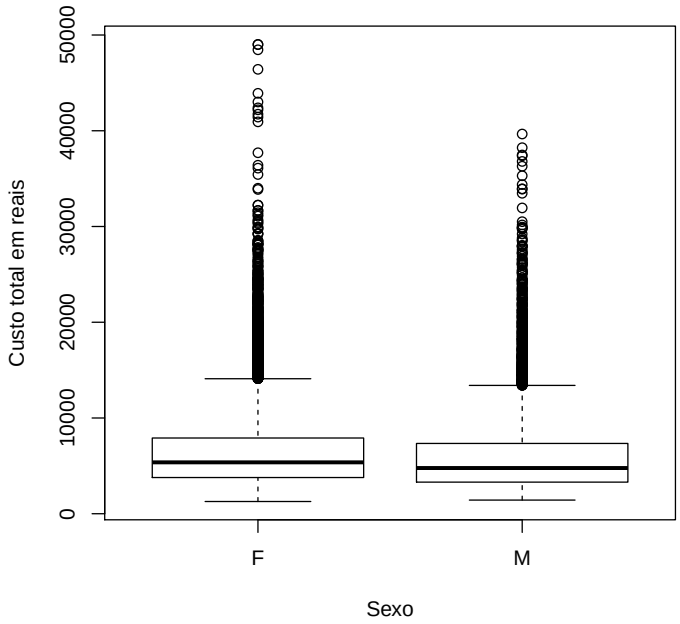


Figura 27: Distribuição do custo total por sexo - Cluster III

As séries temporais contêm até quinze picos e os clientes permanecem, em média, trinta meses em custo maior que zero. Em resumo, as séries oscilam entre meses com custo zero e meses com custos elevados, e o tempo médio de permanência em custo zero é, em média, de sete meses consecutivos. Os picos variam entre 200 e 800 reais.

Os quantis do custo total do grupo III são superiores aos observados para os grupos IV e V, assim como a idade mediana dos clientes. Além disso, as séries de custos referentes ao grupo III também possuem maior quantidade de picos, menor tempo de permanência em custo zero, e picos cujos tamanhos são maiores em relação aos grupos IV e V. Como os clientes permanecem por vários meses com custo maior que zero, a operadora pode intervir neste perfil com programas de promoção à saúde, além do monitoramento desses pacientes, a fim de saber se são realmente doentes ou realizam procedimentos de forma desordenada e sem necessidade.

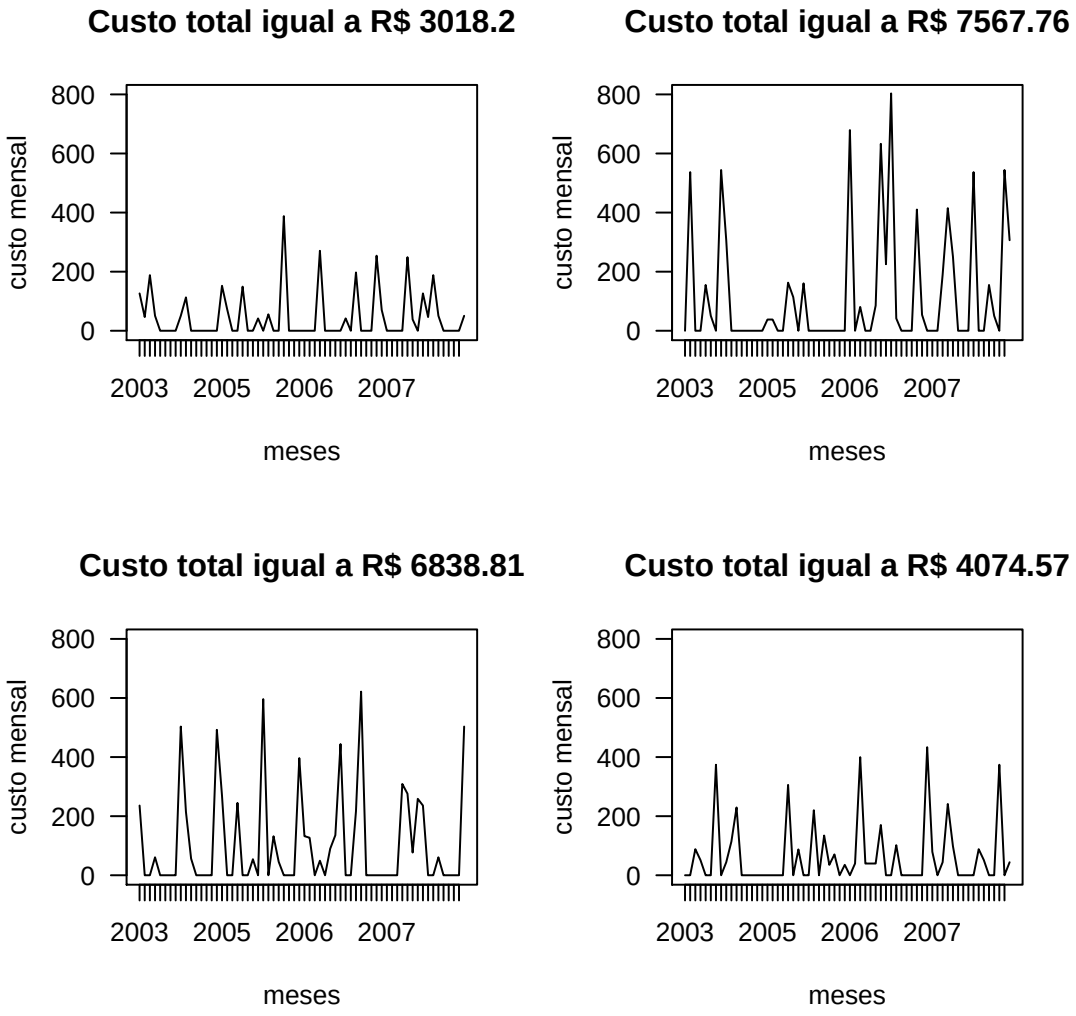


Figura 28: Exemplos de séries de custos - Cluster III

Cluster IV

É composto por 25.285 clientes, os quais representam 25% do total de clientes da carteira. Este grupo é o segundo maior em quantidade de clientes. A proporção de clientes por sexo é 56% e 44% para o sexo feminino e para o sexo masculino, respectivamente. Em média, o grupo gasta 2.292 reais durante os 64 meses de análise e metade dos clientes têm gasto total superior a 1.961 reais.

Tabela 9: Sumário do custo total - Cluster IV

Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
R$479,80	R$1.395,00	R$1.961,00	R$2.292,00	R$2.820,00	R$23.670,00

A idade média dos clientes é cerca de 30,97 anos, sendo 31,81 anos para as mulheres e 29,87

anos para os homens. Vemos, portanto, que a idade média e o custo total médio do grupo IV são superiores ao grupo V. Da mesma forma que observamos no cluster V, as mulheres apresentam idade mediana superior aos homens e a maior porcentagem dos clientes são titulares (cerca de 37%).

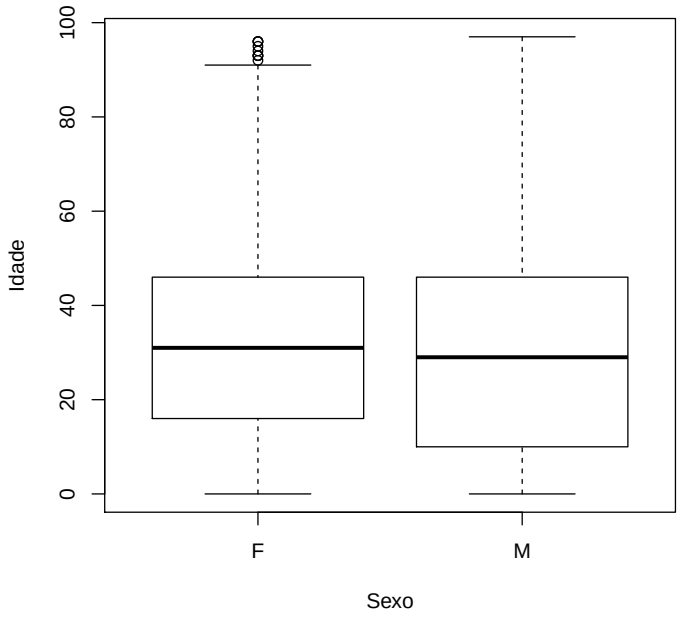


Figura 29: Distribuição etária por sexo - Cluster IV

Quanto à distribuição do custo total, temos clientes que gastam no mínimo 479,80 reais e no máximo 23.670 reais durante os 63 meses de análise. Vejamos a FIG.30: a distribuição do custo total para ambos os sexos apresenta uma pequena assimetria e notamos a presença de outliers. O custo per capita das mulheres (2.359 reais) é maior em comparação ao custo per capita dos homens (2.205 reais).

As séries históricas dos custos são compostas por até 10 picos cujos tamanhos variam entre 100 e 400 reais. Os clientes permanecem, em média, 14,68 meses consecutivos em custo igual a zero e em média, 15,6 meses (não consecutivos) em custo maior que zero. As séries oscilam entre custos iguais a zero e picos (custos acima de 100 reais). Essas são as características desse perfil de clientes no que diz respeito ao custo.

Comparando o cluster IV com o cluster V, o primeiro apresenta maior quantidade de picos, o que resulta em menor tempo de permanência em custo zero. Os picos também são de maior tamanho, o que é um reflexo dos maiores custos totais dos clientes. Então, podemos dizer que,

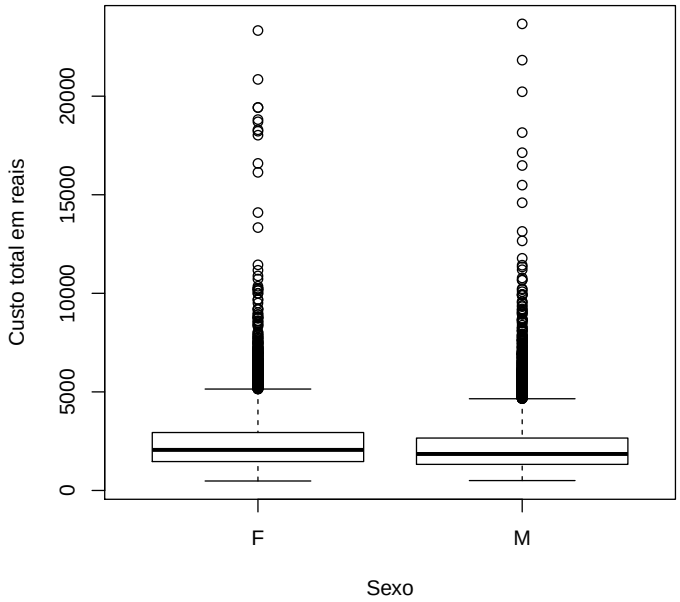


Figura 30: Distribuição do custo total por sexo - Cluster IV

quanto maior o custo médio total, maior o tamanho dos picos. Contudo, os clientes do grupo IV são clientes que realizam mais procedimentos que os clientes do cluster V, mas não realizam procedimentos demasiadamente caros.

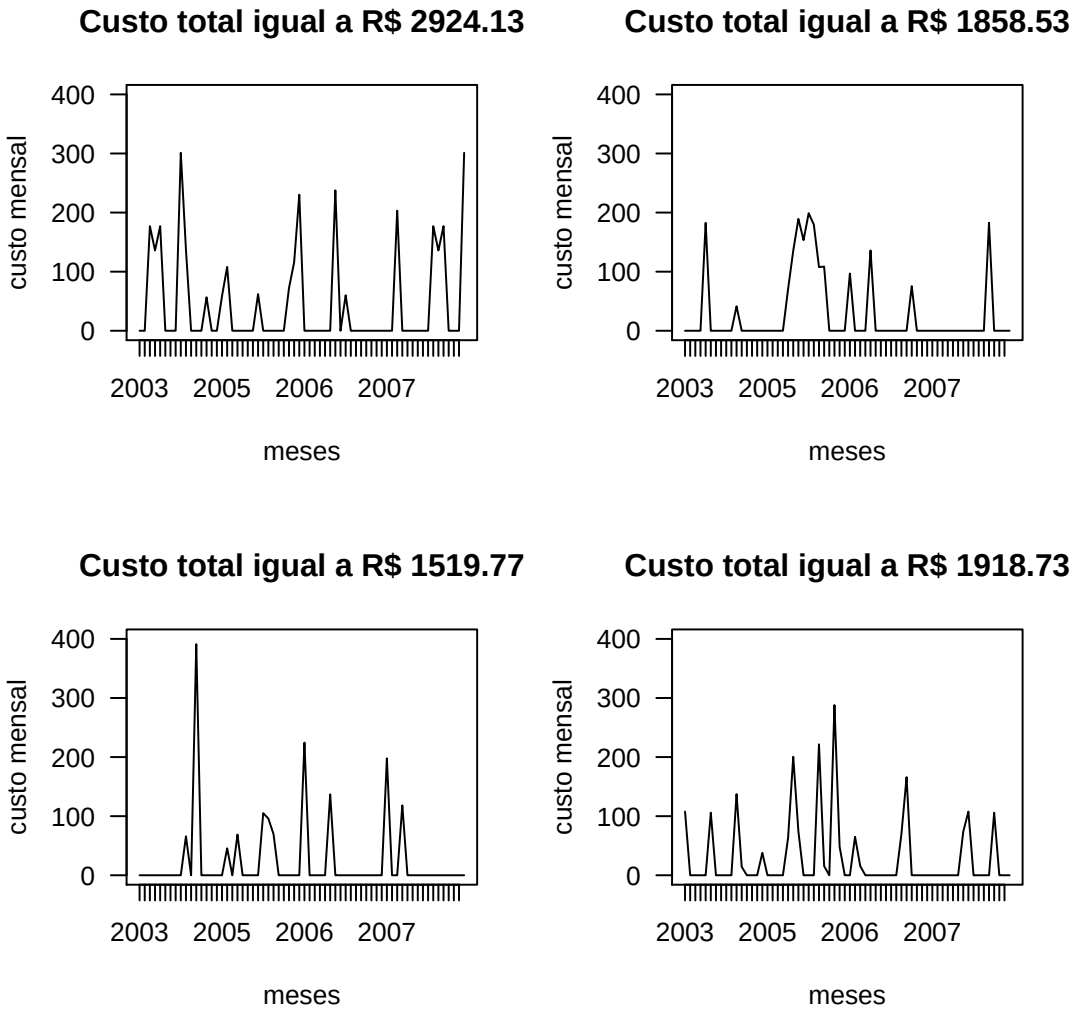


Figura 31: Exemplos de séries de custos - Cluster IV

Cluster V

O cluster V é composto por 14.821 clientes, sendo que 51% desse total são mulheres e 49% são homens. No que tange a titularidade, 37% dos clientes são titulares e 42% são filhos (ou filhas). O custo total médio é 160,8 reais, o custo total mínimo e o custo total máximo são iguais a 3,21 reais e 603,3 reais, respectivamente (TAB.10). Ressalta-se que os clientes do grupo V obtiveram os menores custos totais da carteira.

Tabela 10: Sumário do custo total - Cluster V

Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
R$3,21	R$74,83	R$146,30	R$160,80	R$231,60	R$603,30

Quanto às séries temporais dos custos (FIG.32), observamos poucos picos (no máximo 2),

cujos tamanhos variam entre 50 e 200 reais. Os clientes permanecem em média 46,25 meses consecutivos em custo zero e em média, apenas dois meses (consecutivos ou não) em custo maior que zero. Além disso, as séries de custos contêm o menor tamanho e a menor quantidade de picos e maior tempo de permanência em custo zero em relação a todos os clusters que serão descritos a seguir.

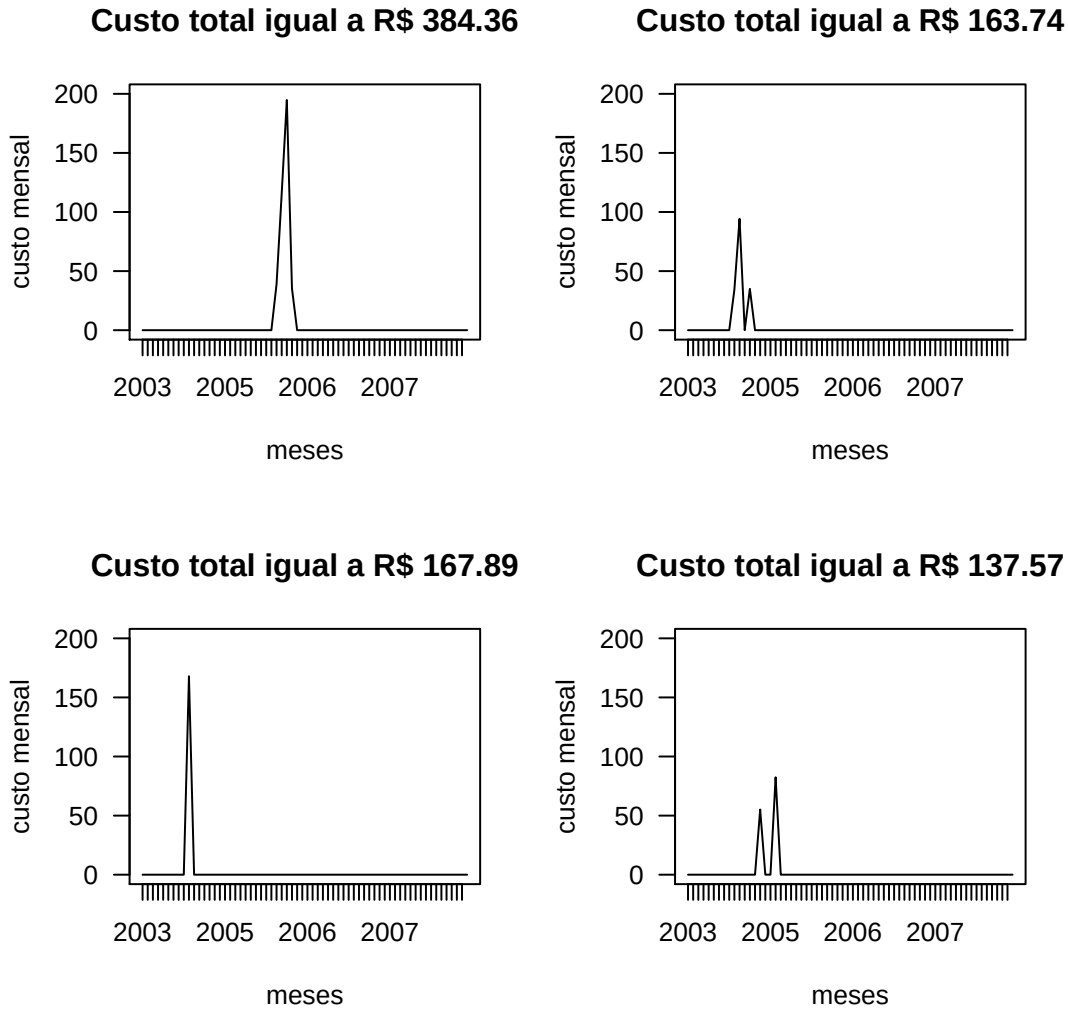


Figura 32: Exemplos de séries de custos - Cluster V

Podemos visualizar a distribuição do custo total através do box-plot abaixo: a distribuição é aproximadamente simétrica e alguns outliers são observados para ambos os sexos. As mulheres têm custo total médio superior aos homens, embora os valores sejam próximos: 164,40 reais para as mulheres e 157,00 reais para os homens.

A idade média dos clientes do grupo V é, aproximadamente, 25,67 anos. Porém, as mulheres

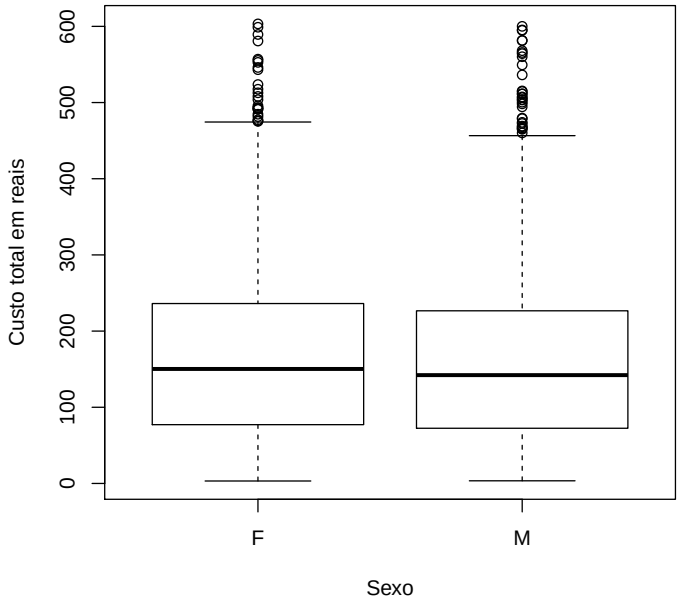


Figura 33: Distribuição do Custo total por sexo - Cluster V

apresentam idade mediana similar aos homens, conforme vemos no box-plot abaixo. A distribuição etária é aproximadamente simétrica para ambos os sexos.

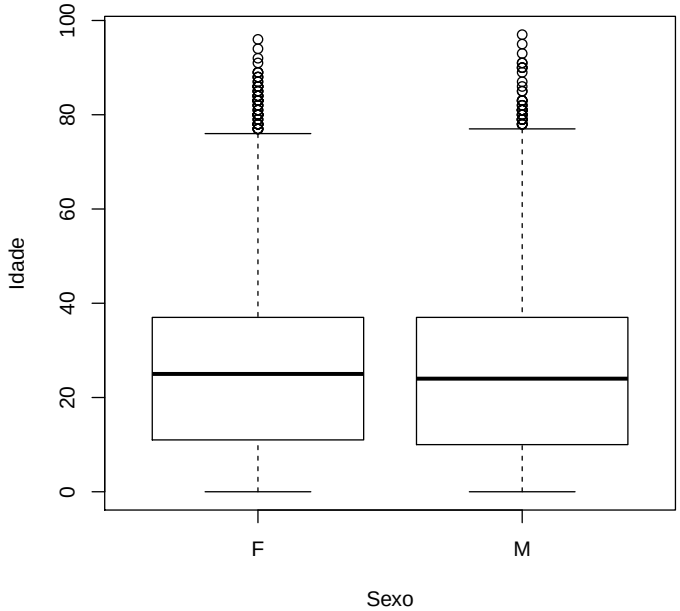


Figura 34: Distribuição etária por sexo - Cluster V

Em geral, o grupo V é formado por clientes jovens, com custos totais muito baixos. Podemos

dizer que são clientes que utilizam o plano esporadicamente, seja quando estão doentes, seja para realizar exames e consultas de rotina.

6 Conclusão

O método de análise de cluster associado às wavelets e ao logaritmo do custo total forneceu grupos de clientes distintos entre si com base nas despesas assistenciais. Sendo assim, atingimos o objetivo proposto: avaliamos a aplicabilidade do método de análise de cluster via wavelets, identificamos os perfis de clientes, assim como as características peculiares de cada um desses perfis.

Com relação aos perfis formados, vemos que quanto maior o custo total médio do grupo, maior a quantidade e o tamanho dos picos presentes nas séries de custos dos clientes. Quanto menor o custo total do grupo, maior é o tempo de permanência em custo zero por meses consecutivos. Em contrapartida, quanto maior o custo total do grupo, menor tende a ser o tempo de permanência em custo zero de forma consecutiva. Ressaltamos apenas o cluster II, o qual obteve custo total médio elevado, mas algumas séries de custos dos clientes são caracterizadas por uma permanência considerável em custo zero ou próximo de zero. Em cada grupo também encontramos alguns clientes com características não semelhantes entre si, o que já era esperado, dado que o método está sujeito a ruídos. Como plotamos aleatoriamente diversas séries de custos em cada grupo, não sabemos quantificar essa proporção de clientes. No que tange à idade, quanto maior o custo total médio do grupo, maior é a idade média do mesmo. Estudos confirmam que pacientes com idade mais avançada consomem uma grande quantidade de serviços de saúde, as internações hospitalares são mais frequentes e o tempo de ocupação do leito é maior devido à multiplicidade de patologias, quando comparado a outras faixas etárias (VERAS, 1994). Tal afirmativa corrobora com o que vemos nos perfis de clientes: há uma correlação entre a idade e o custo dos clientes. A variância é constante em todos os cinco grupos.

É notória as diferenças existentes entre os grupos, principalmente quando comparamos o grupo II e o grupo V. O grupo II possui idade mediana superior a todos os grupos, conforme verificamos no box-plot acima, mas também é o grupo com maior custo per capita. Na FIG.36, no gráfico à direita não plotamos a distribuição do custo total do grupo II devido à sua discrepância, o que causa distorções no gráfico. No gráfico à esquerda, incluimos a distribuição do custo total do grupo II, porém, limitamos o eixo das abscissas a 100.000 reais.

Doutro lado, temos o grupo V com menor custo per capita e composto por clientes mais jovens, os quais não demandam serviços de saúde em grande quantidade. O grupo I e o grupo V possuem características bem similiares: a distribuição do custo total é relativamente próxima

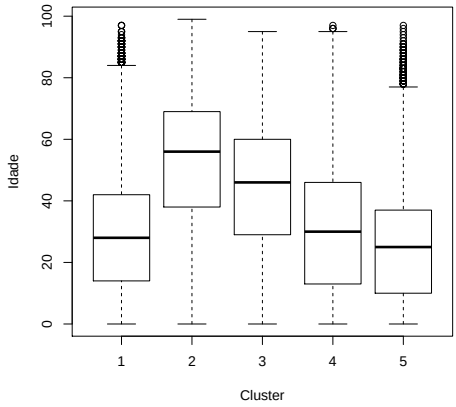


Figura 35: Distribuição etária por cluster

(FIG.36), assim como a idade média de ambos os grupos. O que os diferencia é a quantidade e o tamanho dos picos presentes nas séries de custos, os quais são maiores para o cluster I.

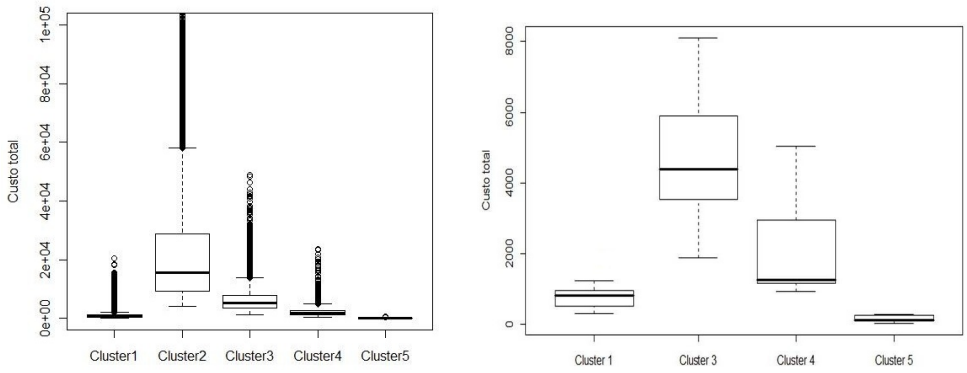


Figura 36: Distribuição do custo total por cluster

A distribuição dos custos totais das mulheres se mostrou similar à distribuição dos custos totais dos homens em todos os grupos. Embora o custo mediano das mulheres foi superior em todos os grupos (exceto no grupo II), tal valor não se apresentou muito discrepante em relação aos homens. O mesmo também foi observado quando comparamos as distrbuições etárias dos homens e das mulheres em cada grupo.

Também observamos a maior proporção de mulheres no grupo II e o maior custo per capita dos homens. Pesquisas realizadas com dados da PNAD apontam as mulheres em idade mais avançada procurando atendimento médico em maior proporção do que os homens, o que pode explicar a maior proporção de mulheres no segundo grupo na carteira como um todo. Assim, elas têm maiores chances de diagnosticar as doenças antes que evoluam, o que aumenta a expectativa de vida das mulheres e melhora a saúde das mesmas. Os homens, por sua vez, não

realizam acompanhamentos médicos periodicamente, o que possibilita o avanço de doenças e gera, posteriormente, gastos exorbitantes para o plano de saúde.

Vemos que o grupo que gerou as maiores despesas para a operadora corresponde ao menor em quantidade de clientes. Tal observação corrobora com a afirmação de que os custos são concentrados e que, um pequeno percentual de clientes do plano de saúde é, de fato, responsável pelos maiores custos.

Em todos os grupos é marcante a proporção de titulares, que pode ser um reflexo da carteira. Nesse caso, não sabemos se os titulares geram, realmente, gastos elevados para a operadora ou em parte de deve aos seus dependentes. Por exemplo, após o parto, o custo do bebê fica atrelado ao custo da mãe, até que ele tenha uma carteirinha. Isso pode elevar os custos dos titulares.

Também encontramos no grupo II clientes com menos de um ano de idade. Isso ocorre, pois, as necessidades em saúde têm um padrão de distribuição, segundo a idade, em "J". Ou seja, as pessoas tanto no início, quanto no final da vida, apresentam mais problemas de saúde. A grande diferença é que as doenças da faixa menor que um ano são agudas e, portanto, de custo menor, enquanto as dos idosos são crônicas e de alto custo.

Segundo Gordilho et al., 2000, em menos de quarenta anos, o Brasil passou de um perfil de morbimortalidade típico de uma população jovem, para um perfil caracterizado por doenças crônicas, próprias das faixas etárias mais avançadas, com custos diretos mais elevados refletidos principalmente no segundo cluster. Essa mudança no perfil epidemiológico acarreta grandes despesas com tratamentos médicos e hospitalares e configura grandes desafios para o sistema de saúde, em especial no que tange à implantação de novos modelos e métodos para o enfrentamento do problema.

Referências

[1]AGÊNCIA NACIONAL DE SAÚDE - ANS. Caderno de Informação da Saúde Suplementar: Beneficiários, Operadoras e planos [Internet]. *Ministério da Saúde*, 2012. Disponível em: http://www.ans.gov.br/images/stories/Materiais_para_pesquisa/ Perfil_setor/Caderno_informacao_saude_suplementar/ 2012_mes03_caderno_informacao.pdf.

[2]AILON, N.;CHARIKAR, M.; NEWMAN, A. Proofs of conjectures in "aggregating inconsistent information: Ranking and clustering" *Technical Report TR-719-05, Princeton University*, 2005.

[3]ASH, A.; ZHAO, Y.; ELLIS, R.; KRAMER, M. Finding Future High-cost Cases: Comparing Prior Cost Versus diagnosis - based methods. *Health Services Research*, 2001.

[4]CEBRIA'N, A.; DENUIT, M.; LAMBERT, M.; ET.AL. Generalized Pareto fit to the society of acturies'large claims database. *North American Acturial Journal*, v. 7, n. 3, 1992.

[5]CUPERTINO, P. Wavelets: uma introdução à teoria, aos algoritmos, e às aplicações. p. 35–94, 2002.

[6]CUPERTINO, P. Wavelets: uma introdução. *Matemática Universitária*, n. 33, p.13–44, 2003.

[7]GORDILHO, A.; SÉRGIO, J.; SILVESTRE, E. J; RAMOS, L. R; ET.AL. Desafios a serem enfrentados no terceiro milênio pelo setor saúde na atenção integral ao idoso. *Rio de Janeiro: UnATI/UERJ*, 2000.

[8]GRAPS, A. An Introduction to Wavelets. *IEEE Computational Science and Engineering*, v. 2, n. 2, 1995.

[9]INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Indicadores Sociodemográficos e de saúde no Brasil [Internet]. 2009 - *[acesso em 15/11/2012]*. Disponível em: http://www.ibge.gov.br/home/estatistica/populacao/indic_sociosaude/2009/com_sobre.pdf.

[10]KAMISHIMA, T.; FUJIKI, J. Clustering orders. *In: Proc. of The 6th Intl Conf. on Discovery Science*, p. 194–207, 2003.

[11]KAMISHIMA, T.; AKAHO, S. Efficient Clustering for Orders. *Mining Complex Data*, v. 165 of Studies in Computacional Intelligence, p. 261–280, 2009.

[12]LEAL, R.; BOAVENTURA, J. B. Perfil etário de beneficiários de planos de saúde de assistência médica no Brasil: uma análise comparativa do mercado individual com o coletivo. *In: Congresso Brasileiro de Ciências sociais e humanas em saúde*, Abrasco, 2007.

[13]LEAL, R.; BOAVENTURA, J. B. Planos de saúde: uma análise dos custos assistenciais e seus componentes. *RAE - revista de administração de empresas*, v. 49, n. 4, 2009.

[14]LIMA, C. R. M; LIMA, C. R. M. A avaliação do custo-eficácia das intervenções em organizações de saúde. *RAE - revista de administração de empresas*, v. 38, n. 2, p. 62–73, 1998.

[15]MAGALHÃES, H. Entrando na Onda *Universidade Federal de Pernambuco*,2001.
Disponível em: http://www2.ee.ufpe.br/codec/slides_ seminario.pdf.

[16]MAGALHÃES, H. Análise de sinais para engenheiros: Uma abordagem via Wavelets *1ª ed. [S.l.]: BRASPORT*, 2007.

[17]MEDICI, A. C; MARQUES, R. M. Sistemas de custo como instrumento de eficiência e qualidade dos serviços de saúde *Caderno Fundap*, n. 19, p. 47–59, 1996.

[18]MEYER, Y. Ondelettes et functions splines *Technical Report, S´eminaire edp, Ecole Polytechenique, Paris, France*, 1986.

[19]MORETTIN, P. A. Ondas e Ondaletas. *1.ed., São Paulo, Edusp*, p.159–180, 1999.

[20]NASON; SILVERMAN. Pacote wavethresh: Wavelets statistics and transforms. *R package version 4.6.1 [Internet]. 2012 - [citado em 20/11/2012].*
Disponível em: http://cran.r-project.org/.

[21]OGDEN, R. T. Essential wavelets for statistical applications and data analysis. *Department of Statistics ,University of South Carolina, Columbia,* p.13–28, 1965.

[22]PHILIPS, R. D. The economics of risk and insurance: a conceptual discussion, in Harold D. Skipper. *Jr., ed., International Risk and Insurance: An Environmental-Managerial Approach (Boston, Ma.: Irwin McGraw-Hill)* p.29–57, 1998.

[23]R Development Core Team. R: A language and environment for statistical computing [Internet]. *Viena; R Foundation for Statistical Computing*: 2008.
Disponível em: http://www.Rproject.org..

[24]VERAS, R. País jovem com cabelos brancos: a saúde do idoso no Brasil. *Rio de Janeiro: Relume Dumará/UERJ*, 1994.

[25]VIDAKOVIC, B.; MULLER, P. Wavelets for Kids *Health Services Research*, 1994. Disponível em: http://www2.isye.gatech.edu/ brani/wp/kidsA.pdf.

7 Anexo A

Tabela 11: Descritiva dos custos dos clientes referentes ao Cluster V

Sumário do custo total - em reais					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
3,21	74,83	146,3	160,8	231,6	603,3

Quantis do custo total - em reais						
0,10%	0,50%	1%	2%	5%	10%	50%
7,7542	17,64	31,84	34,45	34,45	41,34	146,34
70%	75%	80%	90%	95%	99%	100%
213,41	231,6	250,82	300,71	342,72	423,714	603,28

Sumário do logaritmo do custo total - em reais					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
1,166	4,315	4,986	4,843	5,445	6,402

Quantis do logaritmo do custo total - em reais						
0,10%	0,50%	1%	2%	5%	10%	50%
2,047	2,870	3,461	3,540	3,540	3,722	4,986
70%	75%	80%	90%	95%	99%	100%
5,36	5,45	5,52	5,71	5,84	6,05	6,40

Sumário do logaritmo do custo total por sexo - em reais					
FEMININO					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
3,21	77,17	150,30	164,40	236,20	603,30
MASCULINO					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
3,48	72,45	142,20	157,00	226,60	600,10

Tabela 12: Descritiva dos custos dos clientes referentes ao Cluster IV

Sumário do custo total - em reais					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
479,80	1.395,00	1.961,00	2.292,00	2.820,00	23.670,00

Quantis do custo total - em reais						
0,10%	0,50%	1%	2%	5%	10%	50%
583,667	663,0838	711,9368	772,648	902,472	1051,948	1961,15
70%	75%	80%	90%	95%	99%	100%
2595,796	2820,16	3101,626	3945,436	4710,772	6643,794	23668,18

Sumário do logaritmo do custo total - em reais					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
6,173	7,241	7,581	7,604	7,945	10,07

Quantis do logaritmo do custo total - em reais						
0,10%	0,50%	1%	2%	5%	10%	50%
6,37	6,50	6,57	6,65	6,81	6,96	7,58
70%	75%	80%	90%	95%	99%	100%
7,86	7,94	8,04	8,28	8,46	8,80	10,07

Sumário do logaritmo do custo total por sexo - em reais					
FEMININO					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
479,80	1.465,00	2.057,00	2.359,00	2.939,00	23.330,00
MASCULINO					
Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
501,70	1.324,00	1.850,00	2.205,00	2.658,00	23.670,00

Tabela 13: Descritiva dos custos dos clientes referentes ao Cluster III

Sumário do custo total - em reais					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
1.277,00	3.633,00	5.202,00	6.320,00	7.758,00	49.020,00

Quantis do custo total - em reais						
0,10%	**0,50%**	**1%**	**2%**	**5%**	**10%**	**50%**
1.528,42	1.704,70	1.830,20	1.997,19	2.330,34	2.713,59	5202,16
70%	**75%**	**80%**	**90%**	**95%**	**99%**	**100%**
7.052,71	7.758,12	8.590,92	11.290,86	13.998,16	21.313,07	49021,1

Sumário do logaritmo do custo total - em reais					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
7,15	8,20	8,56	8,59	8,96	10,80

Quantis do logaritmo do custo total - em reais						
0,10%	**0,50%**	**1%**	**2%**	**5%**	**10%**	**50%**
7,33	7,44	7,51	7,60	7,75	7,91	8,56
70%	**75%**	**80%**	**90%**	**95%**	**99%**	**100%**
8,86	8,96	9,06	9,33	9,55	9,97	10,80

Sumário do logaritmo do custo total por sexo - em reais					
FEMININO					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
1.277,00	3.782,00	5.367,00	6.429,00	7.909,00	49.020,00
MASCULINO					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
1.426,00	3.302,00	4.773,00	6.056,00	7.343,00	39.660,00

Tabela 14: Descritiva dos custos dos clientes referentes ao Cluster II

Sumário do custo total - em reais					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
4.032,00	9.360,00	15.520,00	28.840,00	28.870,00	1.399.000,00

Quantis do custo total - em reais						
0,10%	**0,50%**	**1%**	**2%**	**5%**	**10%**	**50%**
4.302,14	4.602,38	4.881,76	5.166,70	5.866,05	6.749,92	15.515,03
70%	**75%**	**80%**	**90%**	**95%**	**99%**	**100%**
24.542,57	28.865,38	34.789,58	59.485,30	94.786,50	226.231,75	1.399.317,38

Sumário do logaritmo do custo total - em reais					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
8,302	9,144	9,65	9,792	10,27	14,15

Quantis do logaritmo do custo total - em reais						
0,10%	**0,50%**	**1%**	**2%**	**5%**	**10%**	**50%**
8,37	8,43	8,49	8,55	8,68	8,82	9,65
70%	**75%**	**80%**	**90%**	**95%**	**99%**	**100%**
10,11	10,27	10,46	10,99	11,46	12,33	14,15

Sumário do logaritmo do custo total por sexo - em reais					
FEMININO					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
4.032,00	8.870,00	14.290,00	25.480,00	25.270,00	1.005.000,00
MASCULINO					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
4.044,00	10.370,00	17.980,00	34.710,00	35.660,00	1.399.000,00

Tabela 15: Descritiva dos custos dos clientes referentes ao Cluster I

Sumário do custo total - em reais					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
244,30	505,80	728,10	1.061,00	1.152,00	20.410,00

Quantis do custo total - em reais						
0,10%	**0,50%**	**1%**	**2%**	**5%**	**10%**	**50%**
256,07	271,29	286,90	304,62	342,02	385,30	728,10
70%	**75%**	**80%**	**90%**	**95%**	**99%**	**100%**
1.025,93	1.151,69	1.317,30	2.033,45	2.930,23	5.631,88	20.412,00

Sumário do logaritmo do custo total - em reais					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
5,50	6,23	6,59	6,70	7,05	9,92

Quantis do logaritmo do custo total - em reais						
0,10%	**0,50%**	**1%**	**2%**	**5%**	**10%**	**50%**
5,55	5,60	5,66	5,72	5,83	5,95	6,59
70%	**75%**	**80%**	**90%**	**95%**	**99%**	**100%**
6,93	7,05	7,18	7,62	7,98	8,64	9,92

Sumário do logaritmo do custo total por sexo - em reais					
FEMININO					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
246,70	516,20	760,60	1.114,00	1.242,00	18.460,00
MASCULINO					
Mínimo	**1º quartil**	**Mediana**	**Média**	**3º quartil**	**Máximo**
244,30	490,50	687,30	992,30	1.045,00	20.410,00

8 Anexo B

Sumário dos coeficientes de wavelets e do logaritmo do custo total

Coeficientes	Coef1	Coef2	Coef3	Coef4	Coef5	Coef6	Coef7	Coef8	Coef9	Coef10
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º quartil	0,99	1,40	0,00	2,25	1,03	0,00	0,00	2,32	1,42	0,30
Mediana	2,13	2,66	0,70	3,43	1,94	0,96	0,06	3,39	2,11	1,57
Média	2,76	3,15	1,13	3,72	2,04	1,04	0,43	3,54	2,24	1,53
3º quartil	3,88	4,39	1,71	4,85	2,89	1,61	0,78	4,58	3,19	2,21
Máximo	28,13	22,25	15,80	23,23	16,06	8,28	6,09	15,28	11,91	8,11
Coeficientes	Coef11	Coef12	Coef13	Coef14	Coef15	Coef16	Coef17	Coef18	Coef19	Coef20
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º quartil	0,00	0,00	0,00	0,00	0,00	2,74	2,13	1,86	0,41	0,00
Mediana	1,29	0,41	0,00	0,00	0,00	4,14	2,77	2,39	2,11	1,94
Média	1,07	0,70	0,38	0,14	0,00	3,96	2,88	2,24	1,82	1,52
3º quartil	1,69	1,35	0,62	0,10	0,00	4,97	3,95	2,92	2,61	2,36
Máximo	6,09	4,65	4,12	3,13	1,83	11,33	10,24	8,38	5,96	5,34
Coeficientes	Coef21	Coef22	Coef23	Coef24	Coef25	Coef26	Coef27	Coef28	Coef29	Coef30
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º quartil	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mediana	1,78	0,74	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Média	1,25	1,01	0,77	0,55	0,36	0,22	0,11	0,05	0,01	0,00
3º quartil	2,13	1,96	1,84	1,02	0,51	0,20	0,00	0,00	0,00	0,00
Máximo	5,16	4,73	4,30	4,17	3,55	2,87	2,64	2,51	2,17	0,57
Coeficientes	Coef31	Coef32	Coef33	Coef34	Coef35	Coef36	Coef37	Coef38	Coef39	Coef40
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º quartil	0,00	3,50	3,01	2,70	2,57	0,53	0,00	0,00	0,00	0,00
Mediana	0,00	3,99	3,58	3,31	3,06	2,87	2,74	2,68	2,50	1,09
Média	0,00	4,02	3,34	2,91	2,59	2,31	2,07	1,84	1,62	1,42
3º quartil	0,00	4,44	4,00	3,75	3,54	3,36	3,19	3,00	2,86	2,77
Máximo	0,31	8,60	8,43	7,73	6,74	6,11	5,98	5,75	5,57	5,48
Coeficientes	Coef41	Coef42	Coef43	Coef44	Coef45	Coef46	Coef47	Coef48	Coef49	Coef50
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º quartil	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mediana	0,49	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Média	1,22	1,02	0,84	0,67	0,52	0,39	0,28	0,20	0,14	0,10
3º quartil	2,70	2,61	1,81	1,08	0,72	0,47	0,21	0,00	0,00	0,00
Máximo	4,92	4,91	4,81	4,81	4,47	4,43	4,24	3,41	3,36	3,04
Coeficientes	Coef51	Coef52	Coef53	Coef54	Coef55	Coef56	Coef57	Coef58	Coef59	Coef60
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1º quartil	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mediana	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Média	0,06	0,04	0,03	0,02	0,01	0,00	0,00	0,00	0,00	0,00
3º quartil	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Máximo	2,89	2,86	2,74	2,74	2,28	2,18	1,49	0,78	0,39	0,17

Coeficientes	Coef61	Coef62	Coef63	Logaritmo do custo total
Mínimo	0,00	0,00	0,00	1,17
1º quartil	0,00	0,00	0,00	6,48
Mediana	0,00	0,00	0,00	7,55
Média	0,00	0,00	0,00	7,4
3º quartil	0,00	0,00	0,00	8,42
Máximo	0,05	0,00	0,00	14,15