SEGMENTAÇÃO DO COMPORTAMENTO DE UTILIZADORES DE CARTÃO BANCÁRIO

- Avaliação de Estabilidade -

- Estudo de caso real -

Por
Emanuel Augusto Severino de Matos

Dissertação de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão
2008/2009

Orientado por
Professor Doutor Carlos Soares
Faculdade de Economia – Universidade do Porto

FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Emanuel Augusto S. de Matos, mestrando em Análise de Dados e Sistemas de Apoio à Decisão pela Faculdade de Economia da Universidade do Porto, bacharel em Estatística pela Universidade Federal de São Carlos, com pós graduação em Didática e MBA no Brasil com 17 anos de experiência em Gestão de Empresas.

“ Somente conseguimos chegar ao fim se começarmos....”

Agradecimentos:

Agradeço a Deus em primeiro lugar por dar disponibilidade e serenidade a todos os que em meu caminho me ajudaram e ajudam.Agradeço a meus país, minha noiva, meu primo, meus colegas de turma e a todos os professores do MADSAD, principalmente a meu orientador Prof. Dr. Carlos Soares e ao Prof. Dr. Alípio Jorge, quais não me deixaram esmorecer

Abstrat

The evaluation of cluster´s internal stability is the frucal point of this work. The clusters were extracted from a real database with CCC´s methodology, that is available in the SAS Miner. For this evaluation were used similarity indices and distance between clusters centroids to verify the issues above.

Sumário

1 Introdução 1
1.1 Objetivos da Investigação 1
1.2 Overview 3
2 Segmentação e Clustering 4
2.1 Segmentação 4
2.2 Clustering 7
2.3 Cubic Clustering Criterion 13
2.4 Avaliação de Segmentações em Marketing 14
2.5 Segmentação Efetiva 15
3 Análise da estabilidade de Clusterings 17
3.1 Fonseca e Cardoso 17
3.2 Método Distância entre os centróides 18
3.3 Método Semelhanças entre segmentações 18
4 Avaliação Experimental 19
4.1 Dados da Utilizados 19
4.2 Análise Exploratória dos Dados 20
4.3 Métodos e Avaliação 22
4.4 Aplicativos/Análises 24
4.5 Segmentações 28
4.6 Distância entre centroides 34
5 Conclusões 41
5.1 Geral 41
5.2 Trabalho Futuro 43
6 Bibliografia 44
7 Informações Extras / Apêndice 46
7.1 Sementes 46
7.2 Telas SAS 47
7.3 Tabela das Distâncias 51

1 Introdução

1.1 *Objetivos da Investigação*

Atualmente existe uma demanda muito grande de assertividade no processo de Marketing Direto, as diversidades de necessidade dos ambientes econômicos e de seus mercados fazem das empresas "alvos" ao invés de "flechas" dos consumidores. A segmentação tenta abrandar de forma tecnicista e metodológica este duelo entre as forças. O processo de segmentação tem como pano de fundo uma hipótese que existe dentro de toda a diversidade um ou mais grupos que se alinham, isto é, existem grupos onde internamente conceitos ou necessidades ou anseios são similares. Assim sendo, a busca para maximizar este alinhamento com base nestes grupos homogêneos, podemos chamar de segmentação e identificar dentro do universo pretendido os grupos que são distintos damos o nome de Segmentação de Mercado.

O objetivo da Segmentação de Mercado é buscar uma alternativa entre atender os consumidores individualmente ou colocá-los todos dentro de um único perfil. Utilizando a técnica de segmentação podemos tornar as empresas mais eficientes. (Kotler P. , 2000), pois com esta técnica conseguimos rentabilizar ações de mercado para grupos que tem as mesmas ambições ou os mesmos anseios.

Uma grande utilidade dentro da segmentação é quando conseguirmos avaliar e estruturar os grupos com estabilidade interna. A estabilidade interna pode ser caracterizada como um perfil identificado num grupo, numa amostra e que outras amostras retiradas aleatoriamente não se alterará significantemente este perfil. Este conceito é de grande utilidade para gerenciamento, posicionamento e assertividade de público (Fonseca & Cardoso, 2007).

O objetivo principal deste estudo é avaliar e testar a estabilidade interna dos grupos encontrados.

Como objetivo secundário tratamos de verificar a performance do CCC (Cubic Clustering Criterion) desenvolvido pela SAS que gera "automaticamente" a quantidade de clusters, com base no princípio de minimização dos erros médios quadráticos interno dos grupos.

Desta forma, buscando alternativas de segmentação com base no perfil de consumidores vamos trabalhar com dados reais para tornar mais atrativo os processos que estaremos a utilizar. Os dados utilizados dizem respeito aos tipos de despesa de um conjunto de clientes de uma empresa que necessitava procurar dentro de sua base grupos que pudessem ser acionados de maneira que o retorno pudesse sofrer alguma avaliação.

1.2 *Overview*

Neste trabalho científico buscamos alinhar processos de marketing, computacionais e estatísticos na seleção, busca e avaliação de segmentações e sua estabilidade interna.

Iniciamos definindo conceitos gerais que utilizaremos por todo trabalho, posteriormente tratamos os dados de maneira a perceber seu formato e sua dimensão.

Utilizando clustering hierárquico com o SAS (STATISTICAL ANALISYS SYSTEM) definimos as quantidades de grupos que trabalharemos que se encontra na seção 5.4.3. Experimentalmente mensuramos a estabilidade de cada grupo selecionado relacionando aos trabalhos de Cardoso e Fonseca como podemos ver na seção 4.6..1.1.1 e utilizaremos índices de semelhanças entre os segmentos encontrados como proposto no trabalho de Albatineh ET AL, qual se encontra na seção 4.6..1.1.2.

Proporemos uma vertente de metodologia de avaliação da estabilidade com base no artigo “Supermarket customers segments stabililty” (Fonseca & Cardoso, 2007) e uma vertente de comparação de clusters similares com base no trabalho “On Similarity Indices and Correction for Chance Agreement” (Albatineh & Mihalko, 2006)

2 Segmentação e Clustering

2.1 *Segmentação*

2.1..1 Motivações

A Segmentação encontra-se no meio, entre o MASS Marketing e o Individual, quando trabalhamos com MASS Marketing tratamos de nos dedicar à produção, distribuição e promoção em massa de um produto, podemos citar a Coca Cola que vendia apenas seu refrigerante em garrafas de 200 ml. (Kotler P. , 2000). No contraponto temos o Individual, aquele que se preocupa com a diferença do indivíduo. Podemos exemplificar como um fato feito por um alfaiate com as medidas específicas de seu cliente.

O marketing de segmentação é a busca da personalização do grupo, é o marketing que sabe que existem diferenças individuais mas também sabe a necessidade de incrementar a produção, desta forma busca dar alternativas de MASS marketing a Indivíduos que se comportam de maneira similar. Uma oferta flexível mas não personalizada. Temos uma solução básica e opções que se ajustam a flexibilidade necessária de grupos diferentes, pode-se comprar um carro básico e ajustar kits pré-configurados conforme o interesse do indivíduo. Os kits é que fazem o ajuste dos grupos .

Em termos práticos a Segmentação tem efeitos superiores comparativamente aos dois extremos, pois tratamos grandes volumes de indivíduos agrupados por alguma característica que seja pertinente ao processo devido.

A segmentação de clientes no mundo financeiro é requisito essencial para monitoramento de perfis de clientes, ainda mais atualmente onde existe grande quantidade de dados nas instituições financeiras e de crédito. Transformar estes dados em informação relevante qual fará diferença na tomada de decisão, faz parte da essência da segmentação. Quanto se tem grupos homogêneos e estáveis em sua formação/perfil pode-se tratar de maneira diferenciada, mas não perdendo a característica endógena de cada grupo.

Iniciamos os trabalhos motivados a determinar um processo ou metodologia, que nos conduza a verificar a possibilidade de estabilidade, dentro de cada segmento.

Existem técnicas manuais ou automáticas, estaremos focando uma técnica automática assim sendo se ao aplicarmos esta técnica na mesma base ou num subgrupo este não resultar em clusters com mesmo perfil, a estabilidade interna pode ser considerada frágil assim a confiança no modelo será mínima. A verificação desta técnica teve como motivação os estudos sobre a estabilidade de segmentações (Rebelo, Brito, Soares, Jorge, & Brandão, 2007) , e índices de semelhança (Albatineh & Mihalko, 2006) .

Reproduzimos nos dados à disposição os trabalhos de Fonseca (Fonseca & Cardoso, 2007) e Albatineh (Albatineh & Mihalko, 2006), seus efeitos e as características que encontramos farão parte fundamental do processo metodológico de verificação de estabilidade interna dos segmentos. Trataremos os dados com a metodologia Hierárquica (Ward´s), e como nos requisitos definidos em Marketing (Kotler,1998) , que para serem uteis deverão seguir:

- *Measurability*, seu poder de compra, o tamanho do segmento, seu perfil onde existem problemas quando segmentos muito grande, podemos encontrar segmentos que tenham um tamanho grande o suficiente para uma ação sem um custo elevado.
- *Accessibility*, é possível acessar o segmento, podemos "ir de encontro" a estes clusters, ter acesso, este grupo tende a se concentrar numa determinada região possível de se atingir..
- *Substantiality*, requisito de negócio, tem que ter coerência , "não inventemos carro para pessoas de 4 pés" (Kotler & Armstrong, Principles of Marketing, 1996)
- *Actionability*, requisito de atração, devem atrair o segmento, o grupo deve estar buscando ou querendo e podemos despertar algo que lhes faça querer, uma novidade ou uma reinvenção.
- *Differentiable*, devem ser diferentes entre si, homogêneos internamente .Responderem de forma diferente a um mesmo estimulo, quando entre os grupos (Estabilidade Externa) e de mesma forma internamente (Estabilidade Interna) , assim podemos tratar como estabilidade.

Onde buscamos no trabalho de Carmem (Rebelo M. C., 2006), motivação para dentre as características listadas, tratarmos a Estabilidade Interna como conseqüência da "Differentiable" caracterizada por Kotler.

2.2 *Clustering*

Podemos considerar Clustering sendo um processo estatístico de partição de um universo em grupos. No nosso trabalho estaremos utilizando uma Amostra como Universo e sub-amostras como partições. Estes grupos sendo unidos por uma medida de similaridade. O Clustering tem como função a análise exploratória de dados, em nosso trabalho estaremos utilizando uma abordagem de tomada de decisão. (Jain, A.K.; Murty, M.N.; Flynn, P.J., 1999)

Temos 2 grandes métodos de construção (Stum, 1982), o Método Hierárquico e o Método Não Hierárquico.

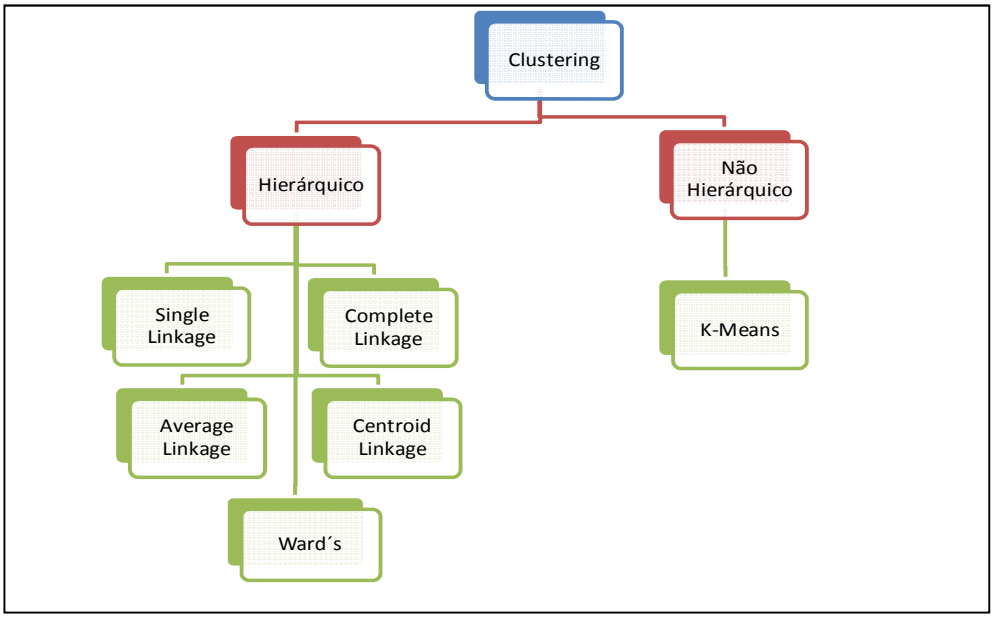


Figura 1 - Clustering

2.2..1 Métodos Hierárquicos

A construção de método hierárquico de similaridade aglomerativa se dá pela junção de indivíduos até o conjunto total, a determinação de número de grupos/clusters é feita à posteriori da formação dos clusters.

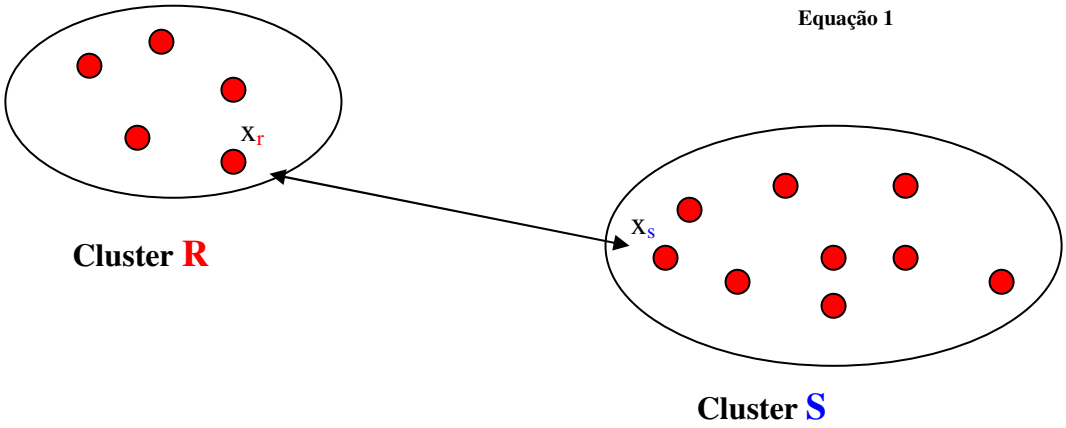
Linkage são critérios de algoritmo de agrupamento quais determinam distância real entre dois conjuntos definindo os dois pontos que representam os conjuntos onde em cada passo do algoritmo hierárquico são agrupados os conjuntos com base no tipo de.Linkage que se está a utilizar.

- **Single Linkage**

Single linkage define a distância entre todos os dois conjuntos como a distância mínima entre eles, isto é a distância entre os dois pontos mais próximos (entidades). Usar este método causa frequentemente o fenômeno de encadeamento, que é uma conseqüência direta do único método do junção que tende a forçar junto os conjuntos devido às únicas entidades que são perto de se não obstante as posições de outras entidades nesse conjunto.

$$d(R,S) = min(dist(x_{ri}, x_{sj}))$$

Equação 1



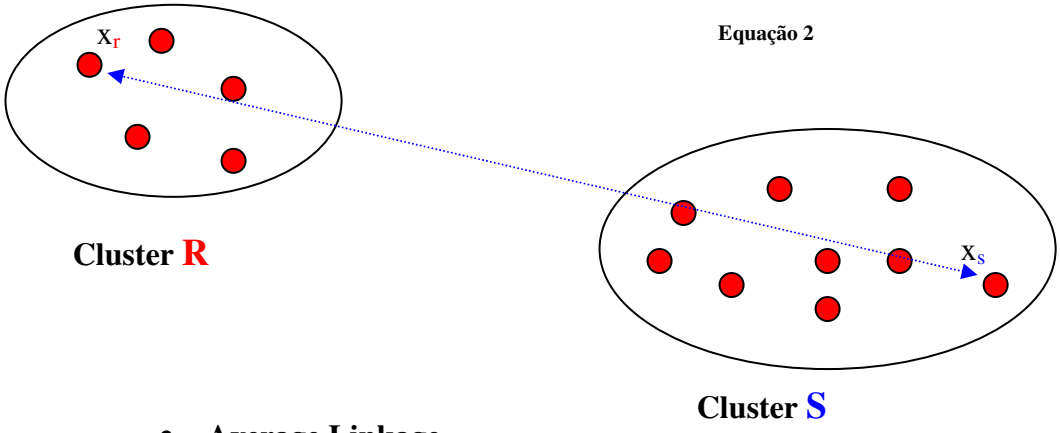
- **Complete Linkage**

Complete linkage é junção que define a distância entre os dois conjuntos como a distância máxima entre eles. Este método não deve ser usado se há muito ruído esperado na série de dados. Igualmente produz conjuntos muito compactos. Este método é útil se

está esperando que entidades do mesmo conjunto serem distantes no espaço multi-dimensional (fornecido não há nenhum ruído, ou seja os outliers são dados mais peso na decisão do conjunto).

$$d(R,S) = max(dist(x_{ri}, x_{sj}))$$

Equação 2

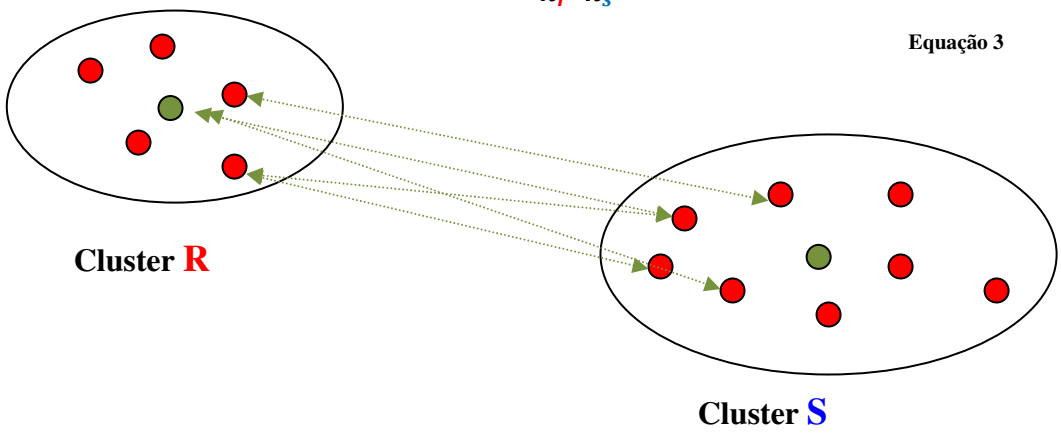


- **Average Linkage**

Average linkage toma a distância média (centro de gravidade) entre todos os pares possíveis de entidades dos dois conjuntos. É conseqüentemente mais computacionalmente cara do que os métodos acima mencionados. Há diversas outras variações deste método, mas se deve compreender que é uma medida intermediária entre o Single e Complete Linkage. O problema de encadeamento não é observado para este método e os outliers não são dados nenhum favor especial na decisão do conjunto, que faz a este método o mais popular dos três.

$$d(R,S) = \frac{1}{n_r \; n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist\left(x_{ri}, x_{sj}\right)$$

Equação 3



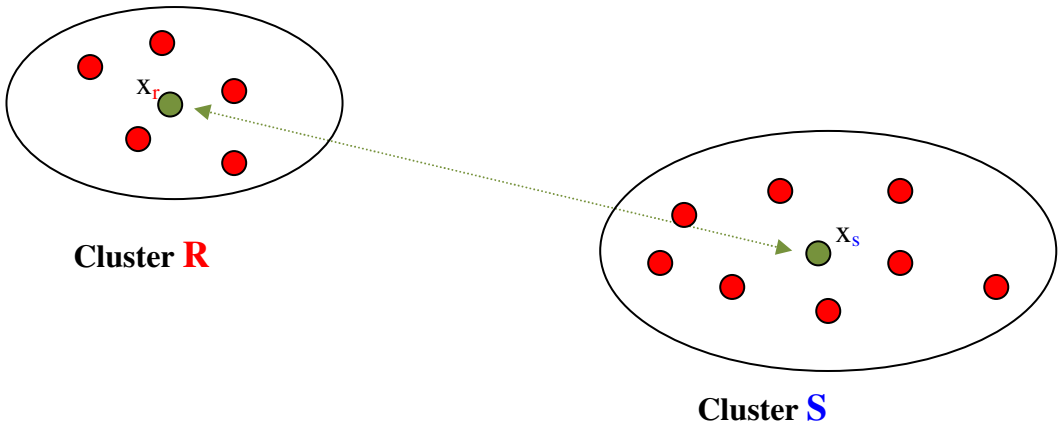
- **Centroid Linkage**

Centroid Linkage utiliza a Distância Euclidiana entre os centros de dois clusters, $d(R,S) = [\![\bar{x}_r - \bar{x}_s]\!]_2$,

Equação 4

Onde :

$\bar{x}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} x_{ri}$, e $\bar{x}_s$ é definido similarmente e $[\![\ ]\!]_2$ é a Distância Euclidiana.



- **Ward´s**

Ward (1963) propôs um procedimento que visa formar agregação das partições P $_{N,}$ P $_{n-1}$,....., P $_{1}$ de um modo que minimiza a perda associada a cada agrupamento, bem como a quantificação do que a perda de uma forma que seja facilmente interpretável. Em cada passo na análise, a união de cada par é considerado possível cluster e os dois pólos cuja fusão resulta em aumento mínimo de 'informações de perda' são combinados. Informações perda é definida por Ward em termos de um erro soma-de-quadrados critério, EES.

$$P_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{\left(\frac{1}{n_K} + \frac{1}{n_L}\right)}$$

Equação 5

2.2..2 Métodos Não – Hierárquicos

Os métodos não hierárquicos são de uma maior facilidade computacional, mas requerem um input à priori que é a definição de numero de clusters a serem formados. Isto é a chave principal do processo dos métodos não Hierárquicos.

- **K-Means**

K-Means é um algoritmo não hierárquico, onde temos que definir à priori (este pode ser o problema) quantos clusters deveremos formar, este algoritmo visa minimizar uma função objetivo, neste caso, um erro quadrado função. A função objetivo :

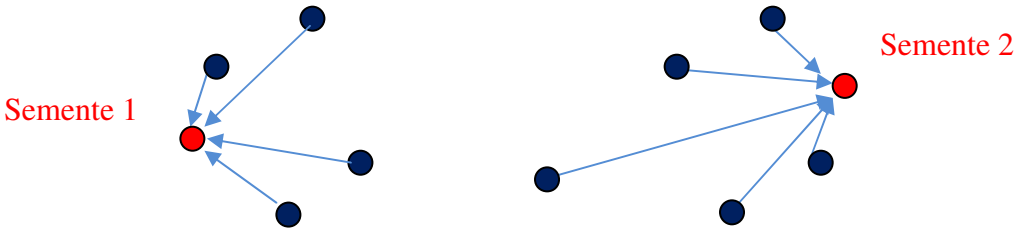
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Equação 6

De onde $\left\| x_i^{(j)} - c_j \right\|^2$ é escolhida uma distância medida entre um ponto dado $x_i^{(j)}$ e o centro do cluster c_j, é um indicador da distância do n pontos dados a partir de seus respectivos centros cluster. A resolução deste algoritmo não significa necessariamente encontrar a melhor solução global. O algoritmo também é significativamente sensíveis ao primeiro cluster que se forma onde os primeiros centros são selecionados aleatoriamente.. O algoritmo pode ser executado várias vezes para reduzir esse efeito.

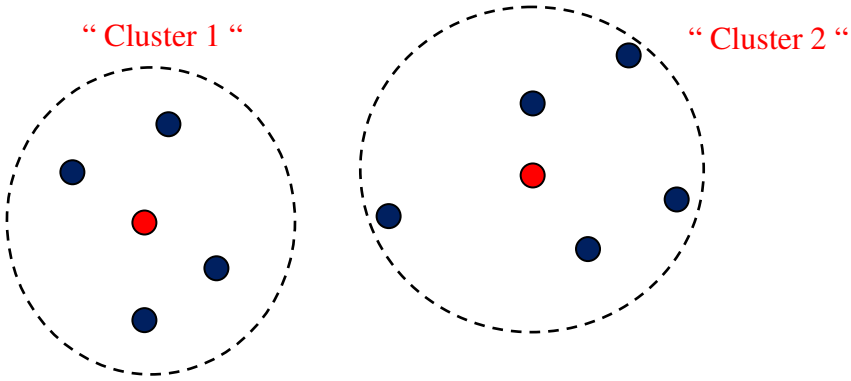
Fase 1 = Sementes aleatórias

Nesta fase são introduzidas sementes aleatórias do numero de clusters desejado.



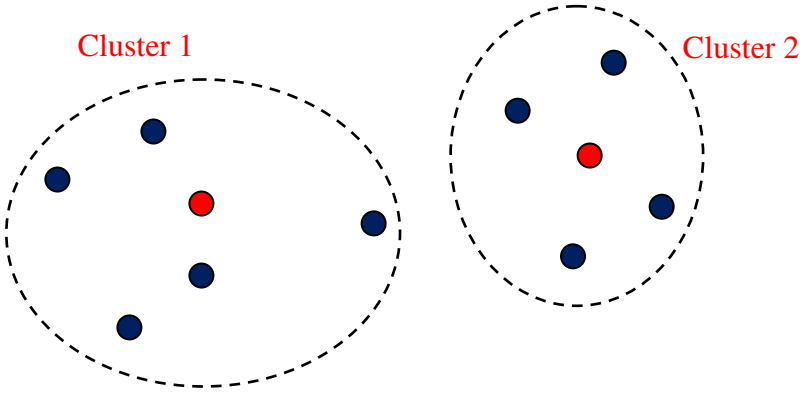
Fase 2 = Sementes centradas

Nesta fase são “centradas” as sementes.



Fase 3 = Criação de novos clusters

Nesta fase se efetiva o resultado da minimização de distancias entre os indivíduos e as sementes, formando no final os clusters esperados.



2.3 *Cubic Clustering Criterion*

Cubic Clustering Criterion - (CCC) foi desenvolvido pela SAS (Sarle, 1983) como uma medida comparativa dos desvios dos agregados da distribuição esperada, se aponta dados que foram obtidos a partir de uma distribuição uniforme. Esta metodologia aponta para o "melhor" numero de grupos que se deve tratar num corte dentro de uma metodologia Hierárquica de Clustering.

Estaremos utilizando e avaliando este critério no formação dos clusters. O critério é calculado como :

$$\mathrm{CCC} = \ln\left[\frac{1 - \mathrm{E}(\mathrm{R}^2)}{1 - \mathrm{R}^2}\right] \mathrm{x}\, \mathrm{K}$$

Equação 7

E quando (R^2) é o esperado R^2, R^2 é observada a R^2 e K é a variância estabilizando-transformação (Sarle, 1983). Maiores valores positivos da CCC indicam uma solução melhor, pois mostra uma grande diferença entre uma distribuição uniforme (sem clusters). No entanto, a CCC podem ser incorrectos se agregam variáveis estão altamente correlacionados.

Calculo do R^2

X = Matriz de dados $n_x p$

$\overline{X}$ = Matriz de média dos Clusters $q_x p$

Z = Matriz indicador de observação, i.é, z_{ik}=1 se a observação i pertence ao cluster k.

Assume-se sem perda que cada variável tem média zero. E Z'Z é uma matriz diagonal que contem n_k (numero de observações no k ésimo cluster) s então

$$\overline{X} = (Z'Z)^{-1}Z'X$$

Equação 8

E o Total da Soma de Quadrados (SST)é

T = X'X

Equação 9

E a Soma de Quadrados entre Clusters (SSBC) é:

$$B= \overline{X}'Z'Z\ \overline{X}$$

Equação 10

E a Soma deQuadrados dentro do Cluster (SSWC) é:

$$W = (X - Z\overline{X})'(X - Z\overline{X})$$

$$= X'X - \overline{X}'Z'Z\ \overline{X}$$

$$= T - B.$$

Equação 11

Podemos mostrar também que o traço de W (trace(W)) é a soma de quadrados da distância Euclidiana de cada observação a média do cluster em questão.

Fazendo T constante, podemos considerar que minimizando o traço de W é equivalente a :

$$R^2 = 1 - \frac{\text{trace}(W)}{\text{trace}(T)},$$

Equação 12

Desta forma chegamos ao R^2. O CCC se obtem por comparação entre o R^2 observado e a aproximação da Esperança de R^2 usando transformação – estabilizada.

2.4 *Avaliação de Segmentações em Marketing*

Em marketing, identificar segmentos se trata de um esforço para aumentar a precisão de acerto de seu publico alvo, conseqüentemente alinhar a estratégia da empresa com sinergia na busca de rentabilizar os esforços coletivos, sempre tendo como balizador o mercado e a industria em que estamos inseridos (Goeller, Susanne; Hogg, Annik; Kalafatis, Stravos P., 2002).

A segmentação em marketing é trabalhada no intuito de identificar grupos a partir de suas preferências, poder de compra, localização geográfica, atitudes de compra e hábitos de compra similares. (Kotler P. , 2000)

Segundo Kotler (Kotler P. , 2000) segmentação em marketing tem como seu ponto de partida como discussão o "marketing de massa", onde todo o esforço é dedicado à produção, distribuição e promoção de um bem para todos os compradores, este conceito se viabiliza desta forma cria-se um mercado potencial.

Este conceito de marketing de massa esta com seus dias contados, pois o marketing de segmento oferece vários benefícios a mais, tanto para o indivíduo como para a empresa, como incrementar a produção de determinado perfil, atendendo um segmento específico, desta forma atuando com foco.

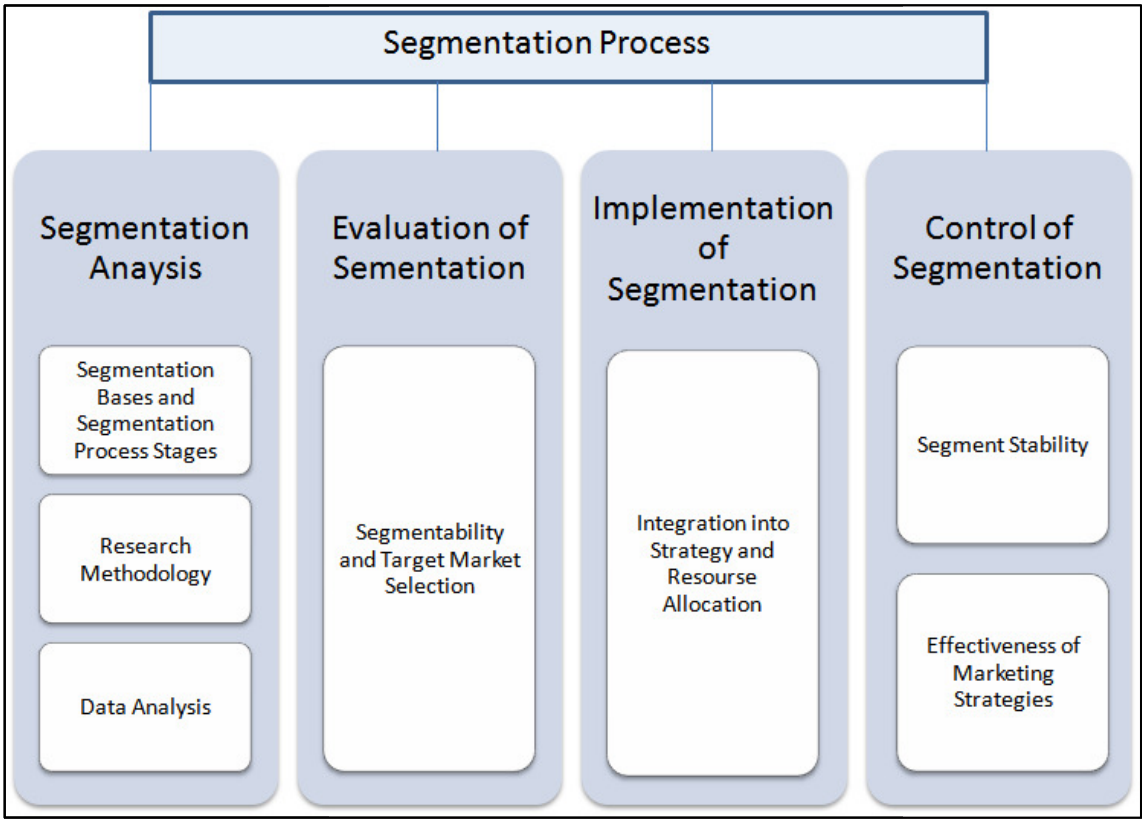
Supõe-se que determinado grupo de indivíduos reagem ou não a determinados estímulos com maior ou menor freqüência, sendo que o que interessa é distinguir dentro de uma população qualquer grupos que tenham atitudes similares a determinados estímulos exógenos.

Podemos identificar segmentos de mercados de diversas maneiras, entre as quais trabalhar três modelos diferentes: Preferências Homogêneas, todos consumidores tem a mesma preferência, o mercado não mostra segmento natural;Preferências Difusas, o outro extremo, a preferência dos consumidores é dispersa por todo espaço e Preferências Conglomeradas, o mercado pode revelar alguns grupos de preferências distintas chamadas de segmentos naturais. (Kotler P. , 2000)

As preferências conglomeradas que serão o foco de nossa análise, isto é, numa determinada população estaremos buscando segmentações naturais.

2.5 *Segmentação Efetiva*

O processo de segmentação pode ser descrito como abaixo, como sugere Goller ET AL em seu artigo " A new research agenda for a Business segmentation":



Segmentation Process

Podemos entender que a avaliação da segmentação se deve a sua eficiência de forma, critério que deve satisfazer a homogeneidade interna do cluster e a heterogeneidade entre cluster, isto é, podemos crer que a assim que junto a este critério, a estabilidade faça parte primordial para um gerenciamento estratégico desta avaliação. (Goeller, Susanne; Hogg, Annik; Kalafatis, Stravos P., 2002)

Nem todas as segmentações são úteis, em marketing para serem úteis devem seguir como visto anteriormente, cap.1.2.

3 Análise da estabilidade de Clusterings

A avaliação dos Clusterings pode ser caracterizada pela procura da minimização das diferenças dentro de grupos e maximização das diferenças entre os grupos com uma ou mais características.

Temos antes que desenvolver o processo de manuseio dos dados, alternando os critérios que possibilitem avaliar de forma eficaz e eficiente a metodologia a ser aplicada na construção dos Clusters, isto dependerá da disponibilidade de ferramentas, complexidade dos dados, tipo de resultado que estaremos buscando.

Podemos tratar com metodologia Hierárquica ou Não-Hierarquica, de maneira aglomerativa ou divisiva, buscando informações exploratórias ou que nos balizem decisões. Questões que temos que retratar para identificar o caminho a percorrer.

A partir dos Clusters formados pode se relacionar três tipos de avaliação, a avaliação externa, comparando com estruturas à priori já definidas; a avaliação interna , verificando se as estruturas internas são apropriadas aos dados e um teste comparando duas estruturas e suas medidas. (Jain, A.K.; Murty, M.N.; Flynn, P.J., 1999). Estaremos utilizando testes para verificar a estabilidade, estes testes foram descritos em Cardoso e Fonseca e índices de similaridade por Albatineh ET AL.

Nosso estudo tem como limitação tratar de uma única Metodologia e métrica assim possibilitando um ponto de avaliação.

3.1 *Fonseca e Cardoso*

A Estabilidade é importante critério de julgamento dentro da segmentação de marketing, quando se obtem uma estrutura estável temos um crescimento na utilidade desta segmentação acrescida de forma relevante, diminuindo a distancia entre o modelo teórico e a pratica. (Fonseca & Cardoso, 2007). Foi discutida no artigo Cardoso e Fonseca , algumas características do desenvolvimento do teste de estabilidade, para nosso estudo adaptamos estas características como por exemplo, no artigo eram “Split” de 60% e 40% e dados em tempos distintos ano de 2000 e de 2003, utilizamos 10 Splits e os dados são num único período.

Avaliar a estabilidade dos grupos é mais do que simplesmente testar sua homogeneidade, trata-se de buscar uma economia em escala quando for necessário em aplicação de pesquisas ou ofertar produtos na busca de respostas, sendo estes grupos estáveis, um subgrupo de tamanho inferior nos reportará com a mesma força e custos menores as respostas solicitadas

3.2 *Método Distância entre os centróides*

Aplicando a distância euclidiana entre os centróides encontrados com a técnica do CCC poderemos ter um critério de avaliação desta metodologia fornecida pelo SAS. Calcularemos e eliminaremos as maiores distâncias e quando já não existirem distâncias únicas, trataremos de reagrupar com todos os pontos anteriores a fim de realinharmos o par.

Desta forma podemos encontrar uma mensuração de eficácia e eficiência da metodologia do CCC com base no nosso experimento.

3.3 *Método Semelhanças entre segmentações*

Calculando indices de semelhanças entre as segmentações encontradas utilizando a metodologia desenvolvida por Albatineh (Albatineh & Mihalko, 2006) em seu paper estaremos verificando as condições de existência de certo grau de similaridade entre as possíveis segmentações encontradas. Desta forma poderemos tentar inferir se temos mesmo algumas segmentações ou na verdade são poucas ou uma só que faz sentido.

Quando utilizamos o método de Semelhança buscamos restringir, ou melhor, eliminar clusters que tenham a mesma ou uma estrutura interna semelhante, clusters criados pelo SAS com o CCC.

4 Avaliação Experimental

No mundo globalizado onde temos grande quantidade de informações e onde o sistema financeiro pode ter seu desenvolvimento acelerado pelo volume de informações que suas bases de dados possuem, e necessário extrair estas informações dos dados, pois as bases por si só não disponibilizam informação.

Desta forma utilizando uma base financeira de dados reais, foi nos proposto o desafio de extrair informação desta base.

Alem de informação, foi nos proposto avaliar um processo de avaliação de segmentação, numa realidade diferente daquela que já dispúnhamos. Assim a partir de agora estaremos entrando num mundo único onde a investigação deverá se mostrarar eficaz e eficiente na construção das propostas acima.

4.1 *Dados da Utilizados*

Foram disponibilizados dados de 5000 clientes, descritos pelas variáveis abaixo.

Escala de idade (Esca_Idade): obtida com base na distribuição da idade dos clientes de forma a que para cada valor exista um número razoável de clientes, onde o 1 significa idades menores e o 5 idades avançadas.

Escala de gastos (Esca_Gastos): obtido por comparação do valor total de gastos com cartões feito pelo cliente com os percentis obtidos com base em todos os clientes. O valor é estabelecido separadamente para os clientes de cada valor da escala de idade da forma similar a Escala de Idade onde o 1 é o de menor percentil e o 5 o de maior.

10 variáveis representando os gastos com cartões (de débito e crédito): CARS, TRAVEL, CLOTHES, HOME, BEAUTY, FOOD, KIDS, EDUCULTURE, HOBBIESFUN, MONEY (esta última representado levantamentos de dinheiro). Assim temos a variavel Esca_Gasto formada composição de outras variáveis, desta forma esta variável não fará parte do trabalho em questão.

4.2 *Análise Exploratória dos Dados*

4.2..1 Variáveis

Abaixo as estatísticas de cada variável, indicando seu mínimo, máximo, média e desvio padrão.

Estatísticas Descritivas

	N	Mínimo	Máximo	Média	Desvio padrão
Esca_Idade	5000	1	5	4,17	,865
CARS	5000	,00	176814,87	549,1378	2677,27980
TRAVEL	5000	0	48100	229,44	1131,585
CLOTHES	5000	,0	20851,5	477,625	1069,9739
HOME	5000	,00	48934,29	531,5187	1281,08851
BEAUTY	5000	,0	24190,2	492,944	1013,6932
FOOD	5000	,00	79385,19	1363,6619	2323,82203
KIDS	5000	,0	13345,9	49,922	350,6587
EDUCULTURE	5000	,0	7293,2	102,321	356,7902
HOBBIESFUN	5000	,0	79635,0	178,488	1409,3762
MONEY	5000	,00	58006,15	414,0715	1641,41373
Valid N (listwise)	5000				

Figura 3 – Estatísticas Descritivas / SPSS

Para cada atributo verificamos com a possível distribuição normal utilizando o SPSS como segue no quadro seguinte.

Testes de Normalidade

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	DF	Sig.	Estatística	df	Sig.
CARS	, 419	5000	, 000	, 100	5000	, 000
TRAVEL	, 420	5000	, 00	, 183	5000	, 000
CLOTHES	, 328	5000	, 000	, 454	5000	, 000
HOME	, 339	5000	, 000	, 385	5000	, 000
BEAUTY	, 313	5000	, 000	, 467	5000	, 000
FOOD	, 279	5000	, 000	, 489	5000	, 000
KIDS	, 443	5000	, 000	, 115	5000	, 000
EDUCULTURE	, 387	5000	, 000	, 302	5000	, 000
HOBBIESFUN	, 450	5000	, 000	, 065	5000	, 000
MONEY	, 400	5000	, 000	, 223	5000	, 000

A. Lilliefors Significance Correction

Figura 4 – Testes de Normalidade / SPSS

Na figura acima, ao nível de 5% de significância não rejeitamos a hipótese da Normalidade nos atributos referidos, pois o nível de significância se encontra em 0

Para que os resultados não sejam afetados pela dimensão absoluta dos valores em cada atributo, estandardizaremos os mesmos.

4.3 *Métodos e Avaliação*

4.3..1 SAS

Utilizamos o Software SAS, que é uma ferramenta utilizada em larga escala no meio profissional e no meio acadêmico, trataremos os dados buscando segmentações que sejam possíveis de mensuração em marketing especialmente quanto à estabilidade que pode-se tratar também por diferenciabilidade, isto é os grupos são homogêneos internamente e conseguimos extrair subgrupos que assim permanecem.

4.3..2 Técnicas de Análise de Dados

Utilizamos como Método Hierárquico de algoritmos de Clustering o Método Ward´s de minimização de variância.

Utilizando o Cubic Clustering Criterion (CCC) [1]que está padronizado no SAS Enterprise Minner e foi concebido por Sarle,1983/SAS, estabelecemos para cada amostra uma quantidade optima de segmentos.

4.3..3 ACP/Análise Multivariada

Faremos o calculo dos Vectores Próprios/Autovectores a fim de reduzirmos a dimensão do polinômio e tentar verificar de forma visual a variância obtida e assim retratarmos os resultados.

4.3..4 Comparação de Segmentação

Estaremos utilizando o Cubic Clustering Criterion (Sarle,1983) como conceito de formação optimo de numero de segmentos por amostra, diferente de Fonseca

[1] Seção 2.3

(Fonseca & Cardoso, 2007) que utiliza o BIC (Bayesian Information Criterion) com variações como metodologia de construção de numero optimos de clusters por amostra.

Com o critério produzido no trabalho de Albatineh (Albatineh & Mihalko, 2006), trataremos de verificar a similaridade das partições dos segmentos obtidos nas amostras que retiraremos do conjunto de dados. O que importa no critério de Albatineh é que tem que existir entre as partições pares de registos que se assemelhem, e neste trabalho apresentam-se 28 indices de similaridade, quais 22 diferentes e onde " são considerados equivalentes e a menor importância é qual índice se deve usar" (Albatineh & Mihalko, 2006).

Desta forma utilizaremos o critério mais usual que é o R encontrado no trabalho de Albatineh (Albatineh & Mihalko, 2006) e desenvolvido na seção 5.4.4.2.2 de nosso trabalho.

D (Distância Euclidiana) calcularemos a distância euclidiana entre os centróides de cada amostra caso geral e/ou seu complementar amostral como sugerido por Fonseca , também calcularemos a **DE** de cada amostra com os centróides obtidos com a totalidade dos dados . Este calculo pretende nos indicar numericamente quanto de distância existe entre os centróides obtidos por CCC, assim não temos a necessidade de trabalhar com um numero idêntico de segmentos por amostra ou obrigar que cada amostra tenha o mesmo numero de partições que o total dos dados, deixando livre para que o CCC defina a melhor partição.

4.3..5 Avaliação de Modelo segundo Marketing

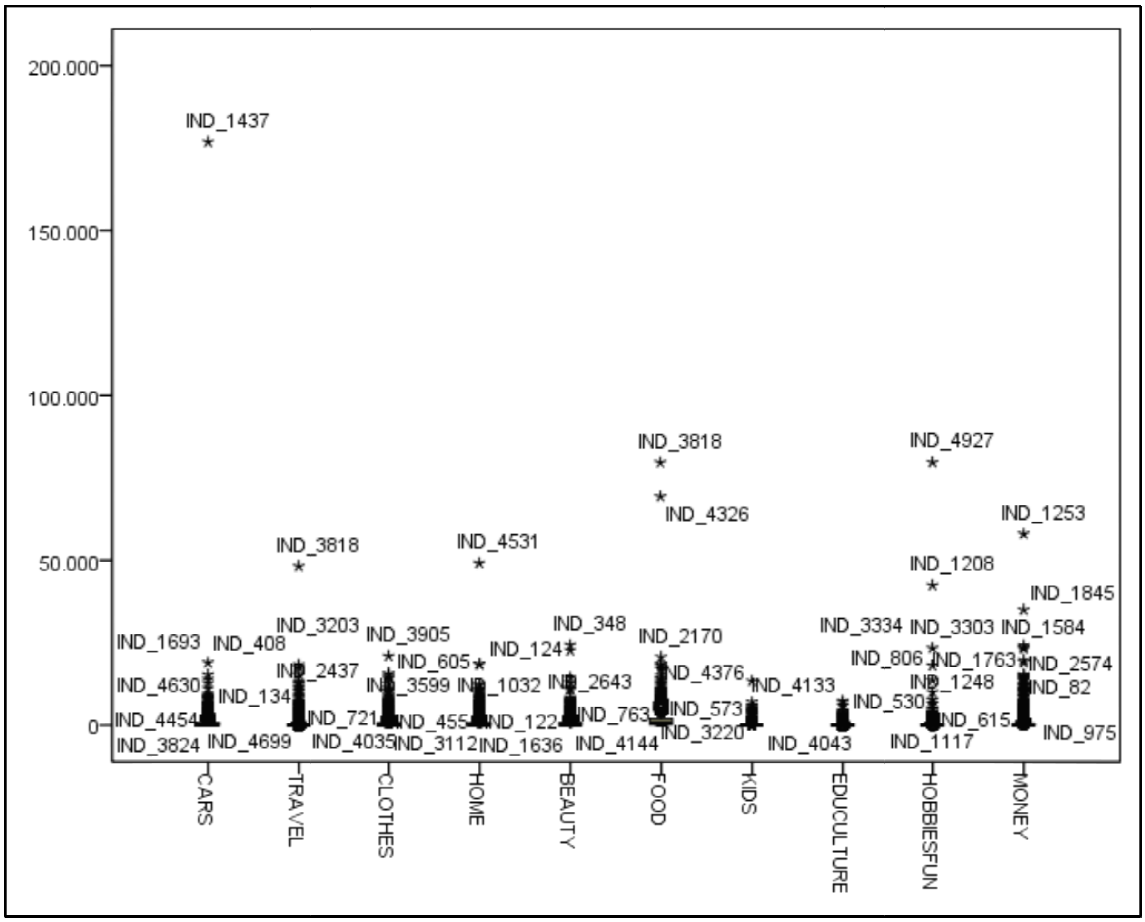
Numa investigação de segmentação dentro do Marketing temos variáveis demográficas , sócio-gráficas e características comportamentais dos potenciais compradores, esta é a definição clássica para um estudo de segmentação em Marketing.

Estaremos avaliando os segmentos segundo sua estabilidade que se revela como sendo um dos critérios de avaliação geralmente usados em Marketing (Kotler & Armstrong, Principles of Marketing, 1996) e de grande importância no gerenciamento do negócio (Fonseca & Cardoso, 2007).

4.4 *Aplicativos/Análises*

4.4..1 Análise Descritiva

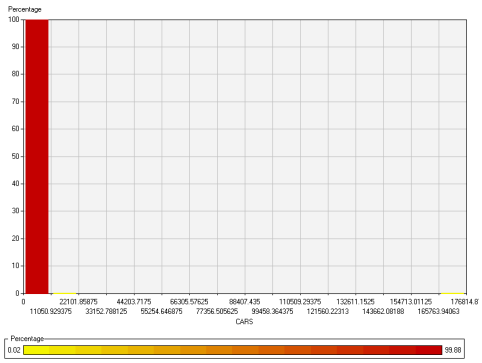
Iniciamos o processo verificando graficamente, com um Box-Plot (Consulting), as variáveis e sua performance, como se segue:

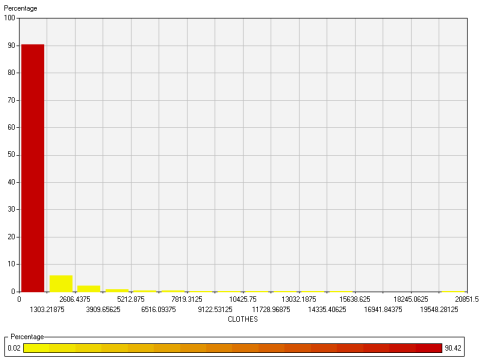


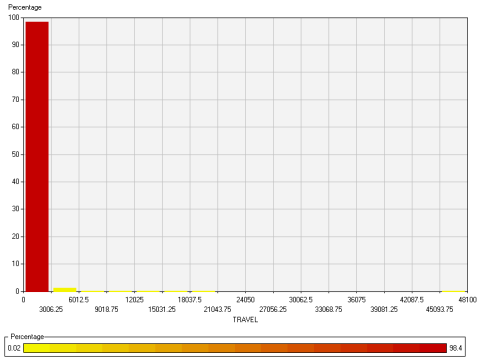
ɡráfica de Outliers/Box -Plot

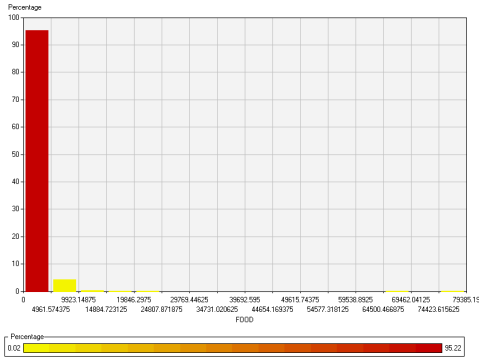
Acima podemos verificar que em todas as variáveis encontramos pelo menos 1 outlier extremo (Maroco, 2007).

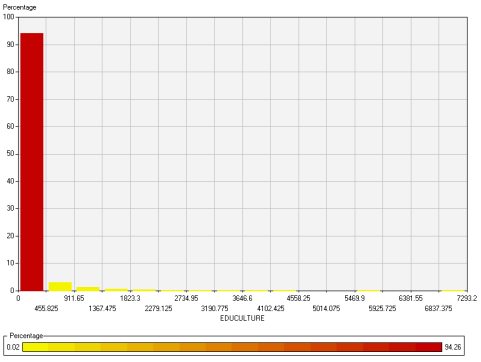
Para cada variável calculamos o seu Histograma, qual está apresentado baixo, notamos que não conseguimos obter informação substancial desta técnica.

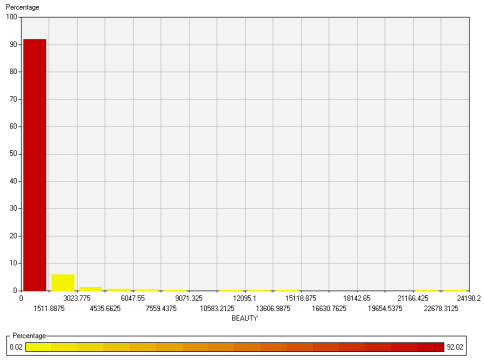


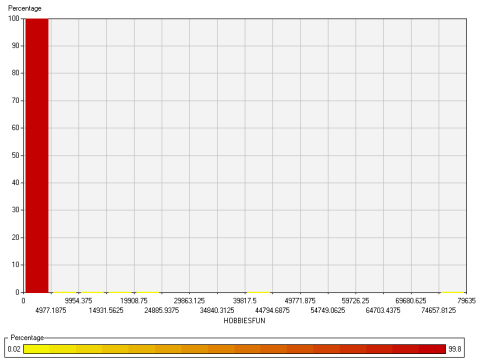


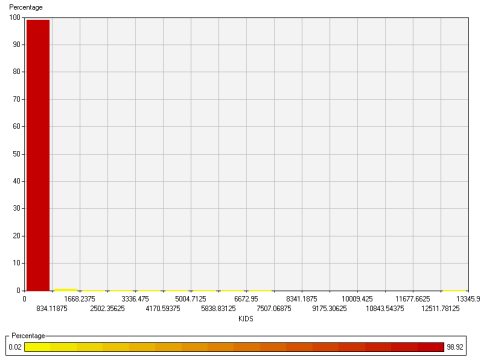


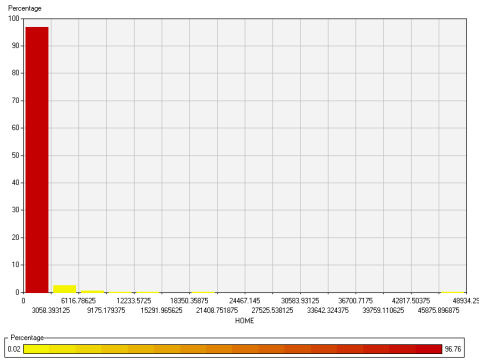


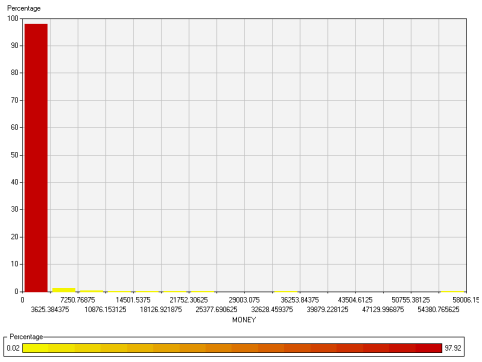












A fim de termos resultados que não levem em conta grandes distorções fizemos um corte de outilers à nível de 0.5% de valores extremos, assim trabalhamos com 4791 registos ao invés dos 5000.

Estandartizaremos as variáveis/atributos para tratarmos todos numa mesma escala de grandeza , excluindo a Escala de Gasto, qual é formada por composição de outras escalas portanto uma combinação de outras variáveis. Utilizando o software SAS e metodologia Ward´s para calculamos as segmentações, e com base no CCC (SAS), o numero de segmentos que estaremos testando . Para complementar a análise calculamos os vectores próprios / autovetores e os componentes principais para utiliza-los graficamente.

4.4..2 Análise Fatorial

Procedemos uma análise fatorial para calcularmos os Vectores Proprios, Componentes Principais e avaliar a variância por autovector. Para o conjunto total dos dados sem os outliers, temos esta análise para podermos tratar de forma visual a comparação entre os segmentos encontrados.

The SAS System

The PRINCOMP Procedure

Covariance Matrix

	KIDS	EDUCULTURE	HOBBIESFUN	MONEY	Esca_Idade
CARS	7541.314	30847.505	45018.254	102389.198	-50.724
TRAVEL	2148.939	20448.829	26669.347	27751.669	-7.187
CLOTHES	27210.969	42847.249	63212.834	84115.369	-46.388
HOME	13694.063	53094.033	61655.417	115669.886	-24.551
BEAUTY	20080.915	38666.306	50571.099	95485.747	14.561
FOOD	41504.209	80042.686	152144.908	247865.720	-77.909
KIDS	13593.457	2211.321	5048.460	4315.951	-10.007
EDUCULTURE	2211.321	52434.977	14035.927	15219.266	-14.818
HOBBIESFUN	5048.460	14035.927	62297.678	23125.924	-23.019
MONEY	4315.951	15219.266	23125.924	679165.297	28.722
Esca_Idade	-10.007	-14.818	-23.019	28.722	0.744

Total Variance 5750703.7759

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2949562.88	2254071.69	0.5129	0.5129
2	695491.19	77733.83	0.1209	0.6338
3	617757.36	127877.86	0.1074	0.7413
4	489879.50	116493.72	0.0852	0.8265
5	373385.78	84809.44	0.0649	0.8914
6	288576.34	57743.10	0.0502	0.9416
7	230833.24	179526.85	0.0401	0.9817
8	51306.39	9409.57	0.0089	0.9906
9	41896.82	29883.26	0.0073	0.9979
10	12013.56	12012.85	0.0021	1.0000
11	0.72		0.0000	1.0000

Figura 4 - Autovalor SAS

Conforme tabela acima, utilizaremos 3 Componentes Principais com a explicação de 74,13% da variância aproximando-se assim do critério de Pearson (Maroco, 2007). Assim podemos explicar a variância dos dados em 3 vertices e suas combinações.

Desta forma os vectores próprios nos fornecem uma redução de dimensão de 10 para somente 3, que ajusta-se a expectativa de visualização e facilidade de tratamento dos dados.

Assim a tabela abaixo nos fornece os dados necessários para redução de dimensão.

The SAS System

The PRINCOMP Procedure

Eigenvectors

	Prin1	Prin2	Prin3
CARS	0.189826	0.075657	-.285889
TRAVEL	0.079717	0.080908	0.062769
CLOTHES	0.250998	0.351815	0.519271
HOME	0.247515	0.410911	0.235639
BEAUTY	0.227386	0.288640	0.354376
FOOD	0.869905	-.425233	-.152885
KIDS	0.018223	0.010505	0.022298
EDUCULTURE	0.038952	0.044166	0.033745
HOBBIESFUN	0.065617	0.029755	0.028991
MONEY	0.136942	0.654237	-.661609
Esca_Idade	-.000031	0.000034	-.000031

Figura 5 - Autovectores SAS

4.5 *Segmentações*

Utilizando o SAS 9.1.3 e a ferramenta Enterprise Miner 4.3 montamos o diagrama abaixo para execução do Métodos de Clustering Hierárquico, Distância de Ward todos os atributos estandardizados, excluindo a Escala de Gastos e outliers. Cada ícone referencia um dispositivo, o primeiro mostra onde buscamos a base de dados, o segundo identifica o filtro de outliers e o terceiro a ferramenta de Clustering que parametrizamos com o Metodo de Ward.

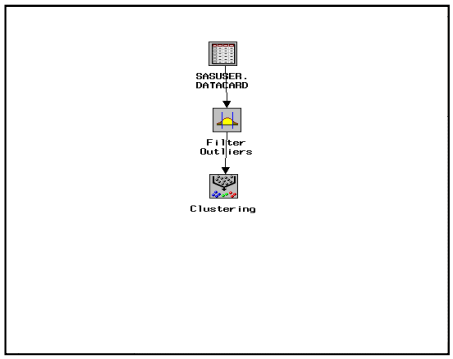


Figura 6 - Estrutura SAS Total

Sequencialmente obtivemos o dendograma, que é o grafo da Análise de Cluster com a Metodologia de Ward, as linhas mais longas mostram a "distancia" necessária para a junção dos Clusters, isto é, quanto maior a linha mais diferentes são os clusters, existe uma diferença maior entre a formação interna dos clusters, diversidade.:

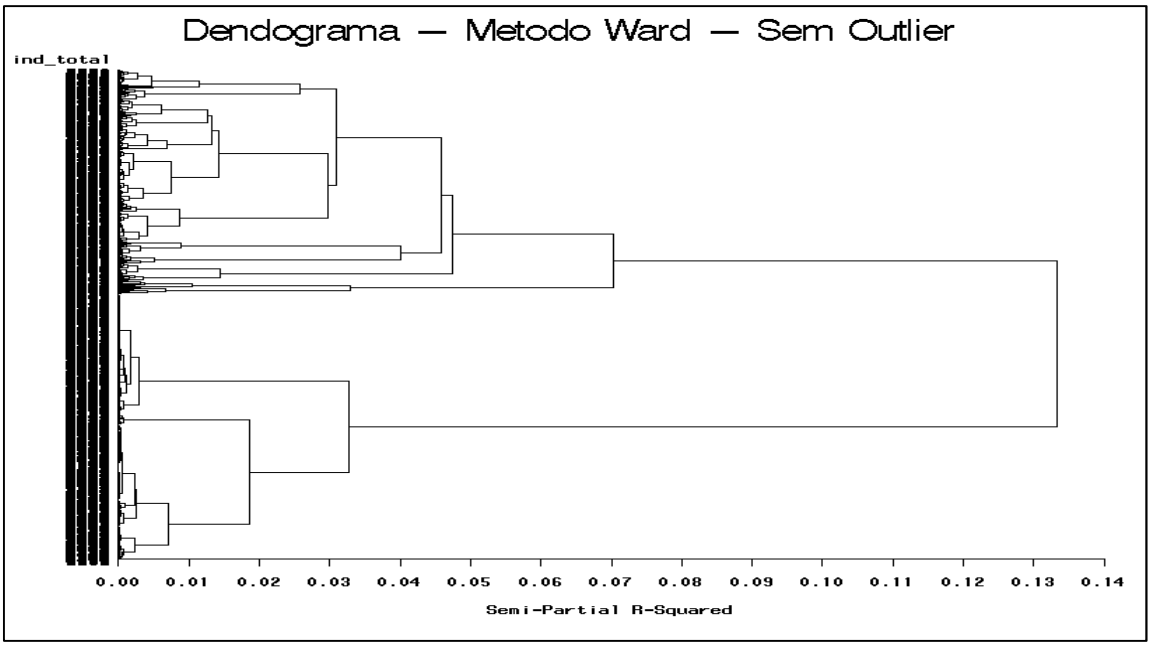
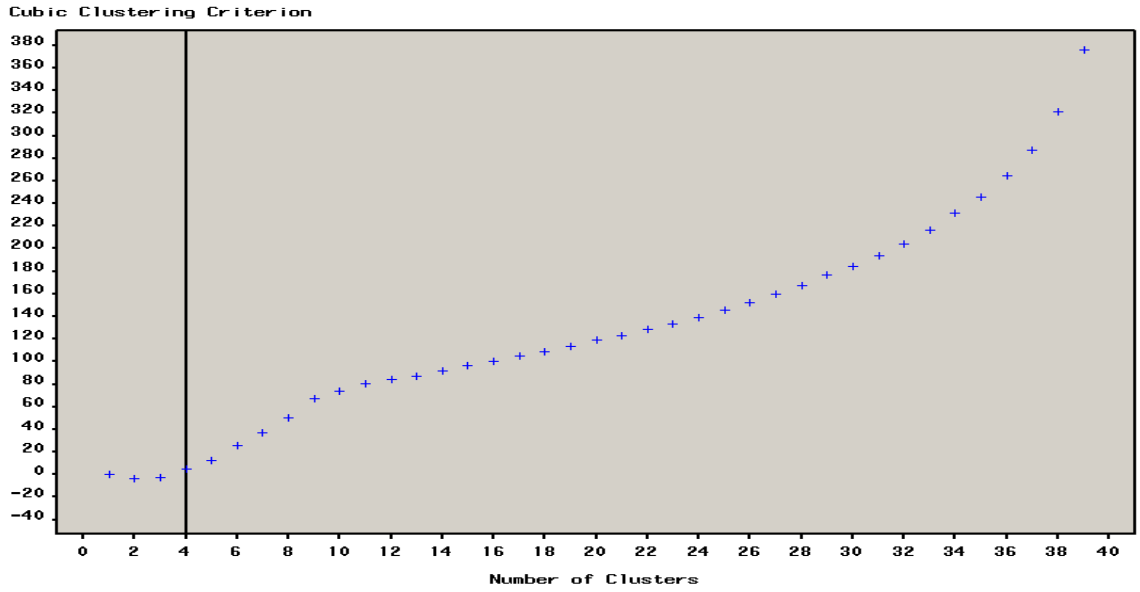


Figura 7 - Dendograma SAS

No Dendograma acima indica a possibilidade de "cortar" em 3 grandes segmentos os dados.

Utilizando a técnica existente no SAS o Cubic Clustering Criterion, com a mesma Métodologia de formação de Clusters, isto é , o Metodo de Ward , esta técnica de obtenção de melhor homogeneidade de Clusters nos indica que 4 segmentos seriam os de melhor ajuste.



stering Criterion - CCC/SAS

4.5..1 Estabilidade Interna

A fim de testar a estabilidade interna, replicamos o processo de CCC para o procedimento treino / teste com 20 amostras de tamanhos 10%, 20%, 30%, 40%, 45%, 55%, 60%, 70%, 80% e 90% assim testaremos a estabilidade interna dos segmentos encontrados com amostras de 10% a 90%.

4.5..1.1 Vectores Proprios

Com base na informação extraída dos Vectores Proprios,utilizando a projeção destes 3 vectores nas variáveis originais, plotamos os gráficos para tentar visualizar qual o comportamento dos clusters encontrados.

Critério Grafico

Utilizando a técnica de ACP para redução de dimensão obtivemos o gráfico a seguir que representa os Centroides encontrados na base completa dos dados[2], com sua localização nos 3 Principais Autovectores / Vectores Proprios:

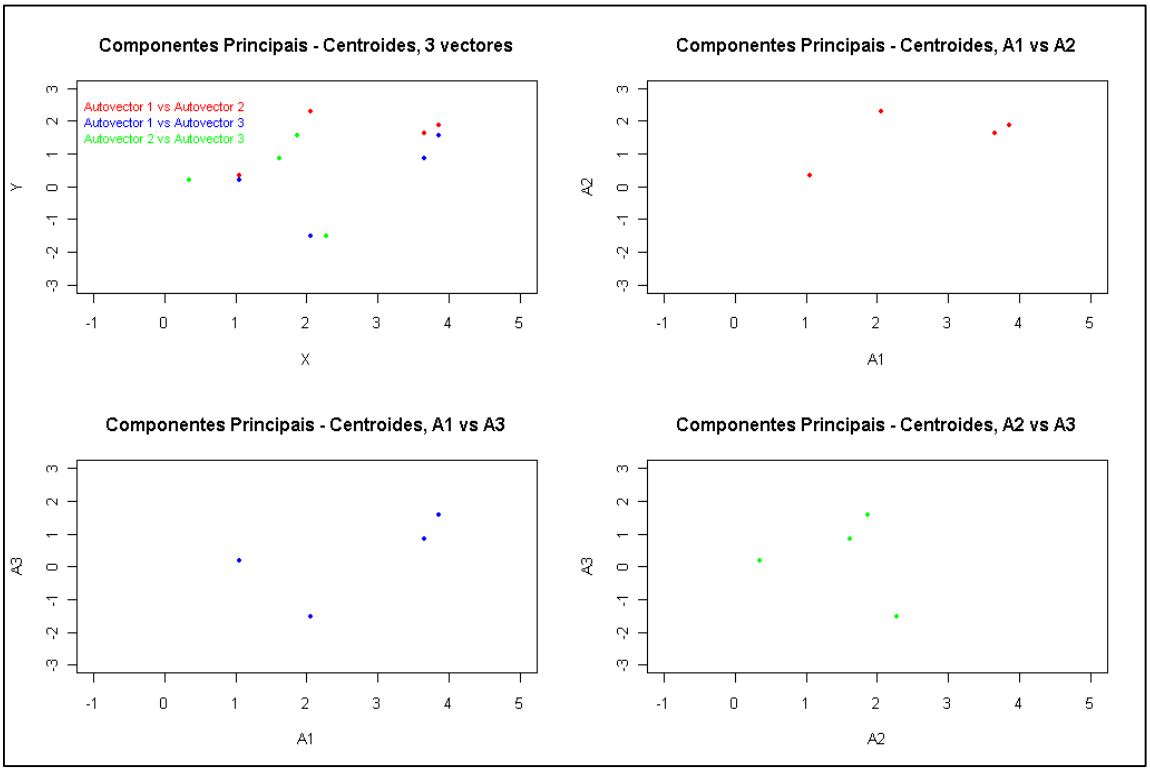


Figura 9 - PCA /Base Total

Para cada % de amostragem projetamos os Autovectores encontrados na PCA da base total, qual resultou nos gráficos à seguir que podemos verificar que quanto menor o % de amostragem , maior é a dispersão da projeção dos centróides, nos induzindo a refletir que o CCC adéqua os centróides na busca da homogeneização interna das partições, assim possibilitando um entendimento de que se mantenham uma certa estabilidade onde a expansão e a contração se faz necessária na adequação dos resultados.

[2] Base excluindo os outlires

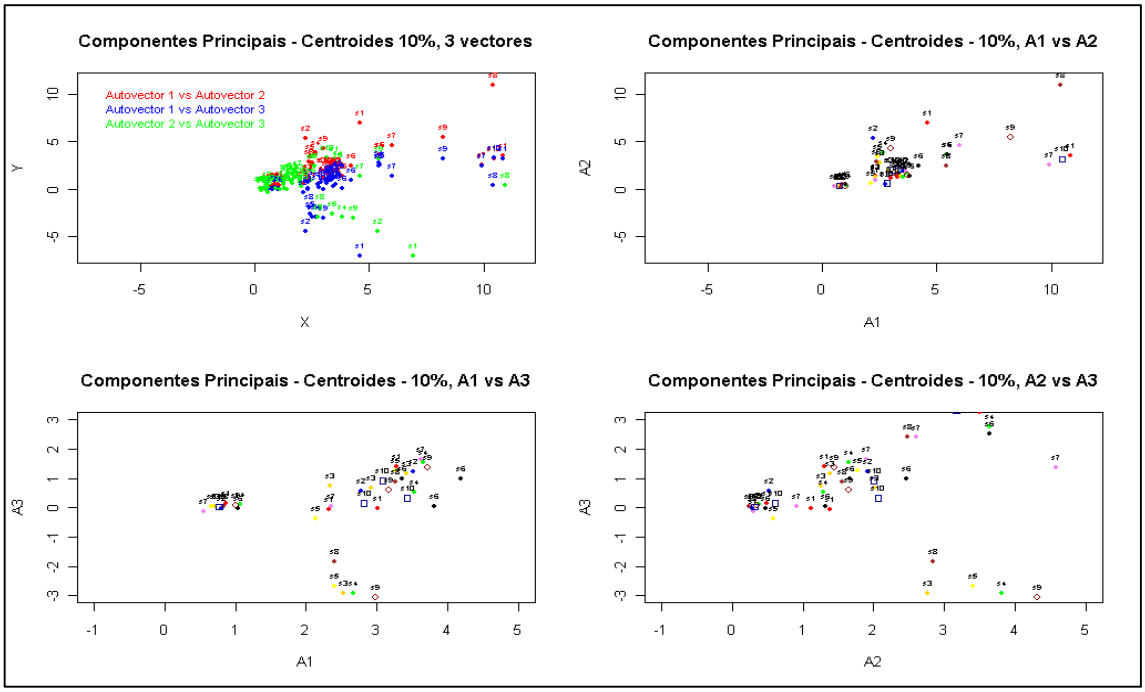


Figura 10 Amostras 10%

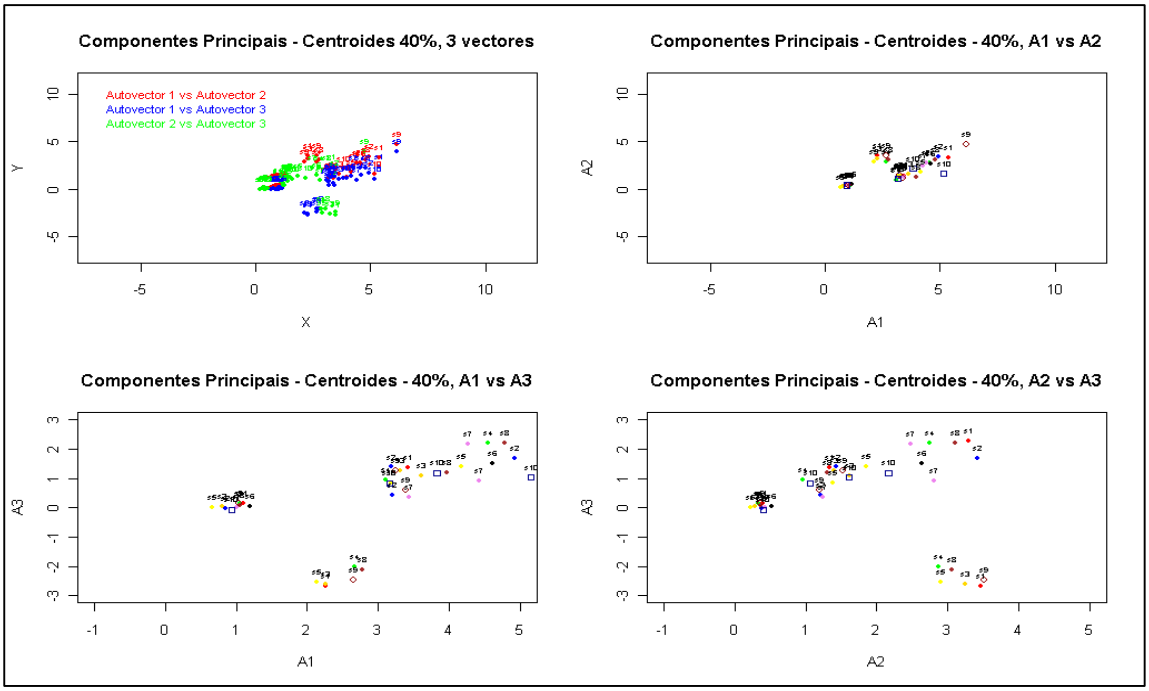


Figura 11 - Amostras 40%

Como podemos notar de 10% para 40% de amostragem vemos uma certa formação de agrupamentos.

Foi feito também para 80% de amostragem os gráficos, onde notamos nitidamente o agrupamento da projeção dos centróides.

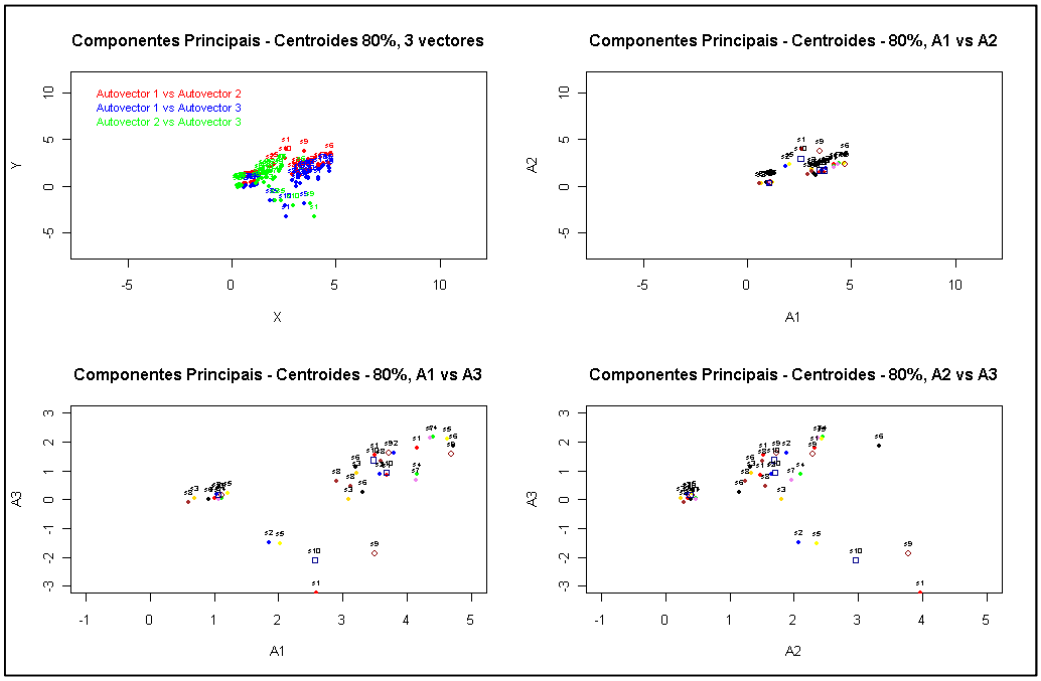


Figura 12 - Amostras 80%

Desta forma foi a experiência utilizando a verificação gráfica das proximidades dos Clusters encontrados.

4.6 *Distância entre centroides*

Replicamos a metodologia de verificação do melhor numero de segmentos encontrados para cada amostra. As amostras foram retiras por uma amostragem simples, e apurado o CCC.

Numero de Segmentos considerados Optimos com o critério CCC

% de Amostragem	Amostras 1	2	3	4	5	6	7	8	9	10	Total
10%	6	4	5	5	4	5	4	5	5	4	
20%	5	4	4	4	4	4	5	5	3	4	
30%	4	3	4	3	3	4	3	4	4	4	
40%	4	4	4	4	4	2	4	4	5	4	
45%	4	5	3	3	4	3	5	3	3	4	
55%	4	4	4	4	4	2	4	4	5	4	
60%	3	2	4	4	3	4	3	3	3	4	
70%	4	4	4	4	4	4	3	5	3	3	
80%	5	4	3	3	3	4	3	4	4	4	
90%	3	5	4	4	3	4	4	3	3	4	
100%											4

Figura 13 - % de Amostras e Total - n. de segmentos CCC/SAS

Acima a tabela que nos refere o percentual de amostragem aleatória simples por numero da amostra , assim para a amostra 1 com 10% de dados isto é dos 4791, algo em torno dos 10% o numero de clusters que o CCC apontou como sendo optimo foram 6, nesta mesma amostra com 80% dos valores o CCC apontou como sendo 5 o numero de clusters optimos. Como o total nos levou a 4 segmentos e as amostras oscilaram entre 3 e 6 segmentos, aplicamos a verificação da Distancia Euclidiana entre os centróides para cada nível de amostra e contra o total.

Calculamos a Distancia entre os centróides obtidos para cada amostra mantendo fixo o percentual amostrado, assim, comparamos as amostras entre si e ao Total obtido. Como notação temos S a amostra principal excluídos os 0.5% de outilers. S1 é a primeira sub-amostra de S, S2 é a segunda e assim por diante até S10 que é a décima sub-amostra. S1(10%) significa que foi retirada a primeira sub-amostra de tamanho 10% com base na Amostragem Aleatória Simples., assim por diante até S10(90%) que configura a sub-amostra S10 de tamanho 90%.

Após retiramos as sub-amostras todas, multiplicamos cada sub-amostra retirada dentro de seu respectivo tamanho, assim S1(10%) x S2(10%), S1(10%) x S3(10%) e até

S9(90%) x S10(90%), este calculo multiplicativo teve como o índice da Distancia Euclidiana, assim temos uma medida informativa de distância de comparação entre os vários centroides encontrados para as amostras retiradas.

Após este calculo fomos eliminando as menores distâncias entre os valores obtidos e deixamos a maior distancia encontrada para cada par (S^1 x S^+)^3, depois calculamos a média dentro de cada Amostra e a sua média de distancia para o Total utilizando os pares que se formaram e as suas respectivas distâncias.

Desta forma verificamos o seguinte gráfico:

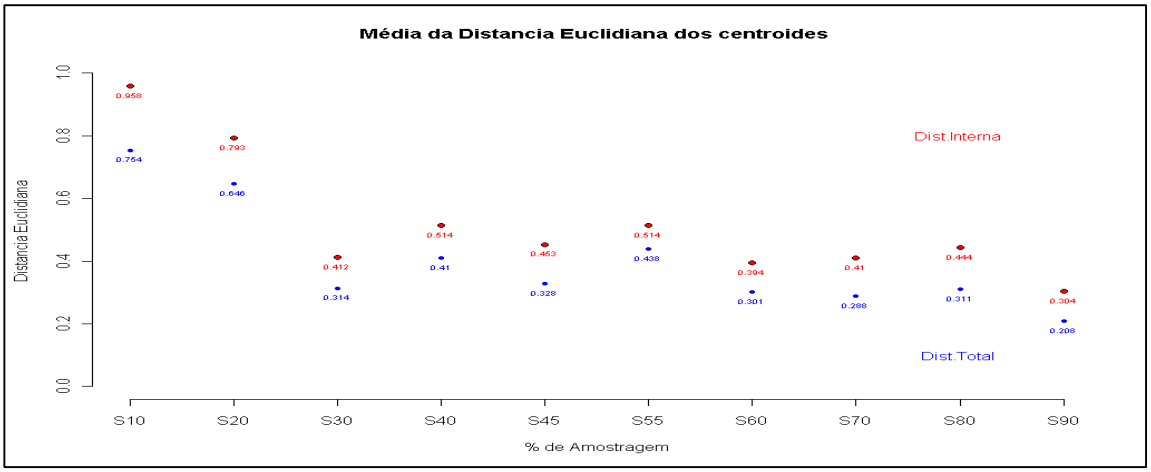


Figura 14 - Distância Euclidiana

Como podemos notar as distâncias entre os centroides das amostras de 10% e 20% são superiores às dos demais incluindo a distância contra o total.

4.6..1.1.1 Metodologia Fonseca

Agora utilizando a metodologia do trabalho Fonseca ET. AL refizemos os cálculos fazendo um split de amostras, isto é quando obtivemos 10% os outros 90% se transformaram em outra amostra e assim comparamos também as distancias destes centroides, obtivemos o gráfico a seguir.

[3] Par de amostras S1xS2 ou S1xS3... correspondente a cada % retido da Amostra Principal (S)

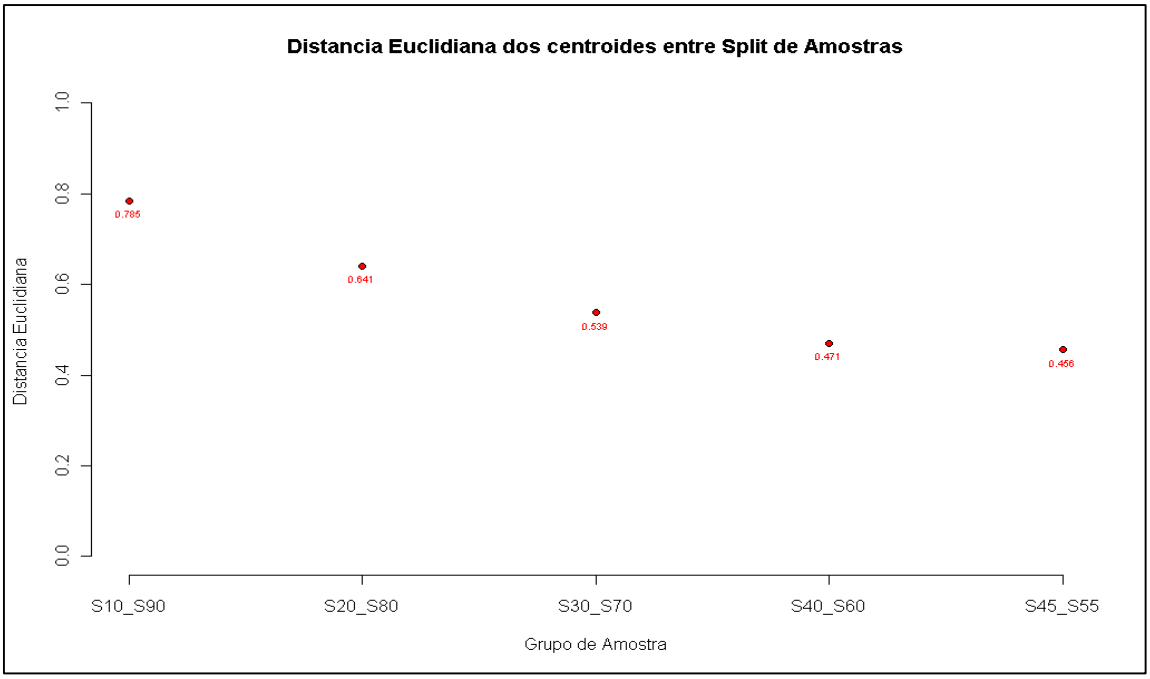


Figura 15 - Distância Euclidiana SPLIT

Temos da forma similar que as distancias entre centróides de amostras com % muito distantes são superiores com relação a amostras onde a partição é mais próxima, isto é amostras de 90% com seu complementar de 10% tem uma média de distância entre os centróides superior a amostras de 55% com seu complementar de 45% .

Todos as amostras seguem o padrão do CCC como optimizador de quantidade de partições.

4.6..1.1.2 Metodologia Albatineh- Similaridade

Utilizando como base o trabalho de Albatineh, calculamos a similaridade entre as partições encontradas nas amostras e o total, esta similaridade leva em conta o par de registos e sua efetiva utilização no segmento encontrado, se um registo A está na S10_1 e não na S10_2, não fará parte do calculo para determinar a similaridade entre os segmentos, portanto para as amostras de 10% utilizamos na comparação somente os

indivídusos que pertenciam simultaneamente a duas amostragem e assim foi realizado para cada % de amostra. Comparamos também a similaridade de cada % de amostra com o Total aferido.

Temos abaixo gráficos comparando a similaridade encontrada comparando os SPLIT[s] das amostras, neste caso foram utilizadas 20 amostras para cada SPLIT e o índice R (Albatineh & Mihalko, 2006) que é o índice de Sokal and Michener(1958) ,Rand(1971) que tem a seguinte equação

$$R = \frac{a+b}{a+b+c+d}$$ **Equação 13**

,onde

Tabela de Similaridade (Albatineh & Mihalko, 2006) para dois Métodos de Clustering,

	Método 2 ou Entrada 2			
Metodo 1 ou Entrada 1	Numero de Pares	No Mesmo Cluster	Em diferentes Cluster	Total
	No mesmo Cluster	a	b	$a+b$
	Em diferentes Cluster	c	d	$c+d$
	Total	a+c	$b+d$	M

E este índice tem intervalo de [0,1] onde 1 é o ajuste perfeito. Utilizando esta metodologia a adaptamos para compararmos dois a dois clusters de nosso experimento. O detalhamento dos índices estão no artigo " On Similarity Indices and Correction for Chance Agreement" (Albatineh & Mihalko, 2006)

Desta forma obtivemos os gráficos a seguir:

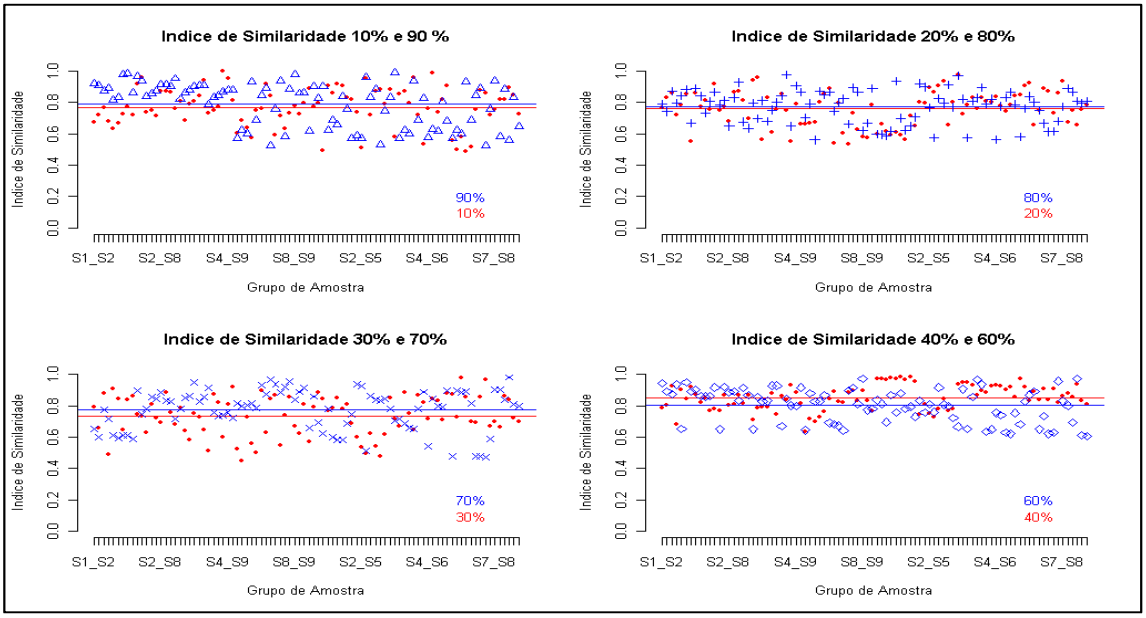


Figura 16 - Similaridade Geral

Assim buscamos fundir dois processos, de Fonseca que sugere o SPLIT dos dados e o treino teste e de Albatineh que busca a similaridade entre partições de segmento, um ajuste que se fez necessário foi não se ater a quantidade de segmentos obtidos em cada amostra e o Total, assim compararmos a similaridade com base na segmentação optima obtida com o CCC.

No trabalho de Fonseca é sugerido a quebra ou SPLIT de 55 e 45 % qual está no gráfico abaixo:

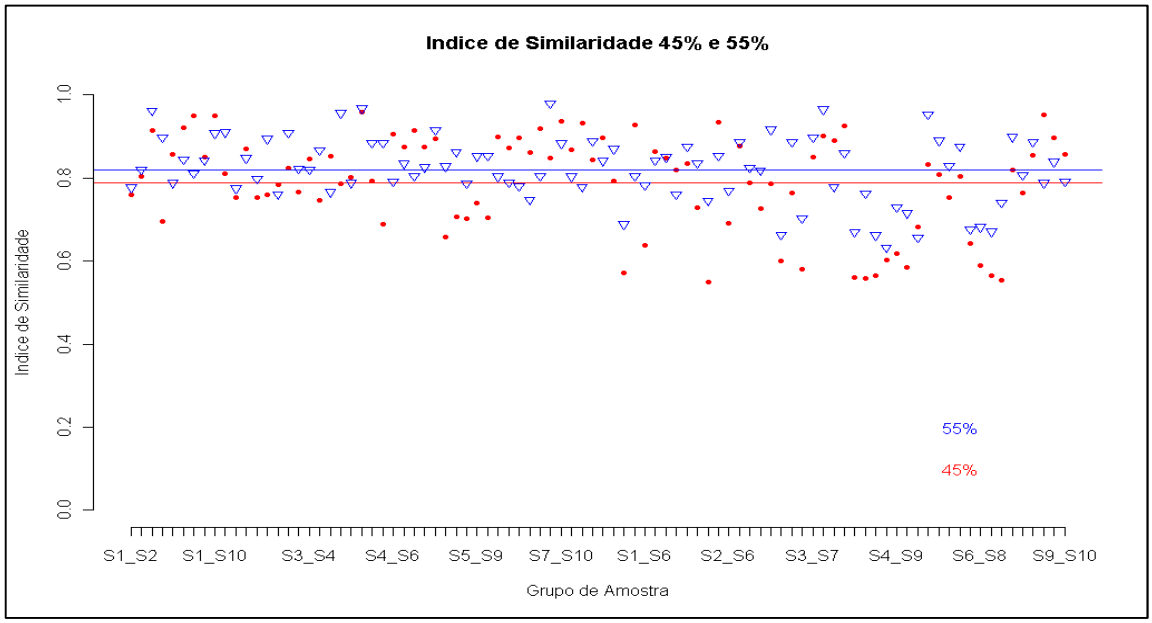


Figura 17 - Similaridade "Fonseca et al"

Comparando os índices de similaridades médio das 20 amostras por amostra e contra o Total, obtivemos o seguinte gráfico:

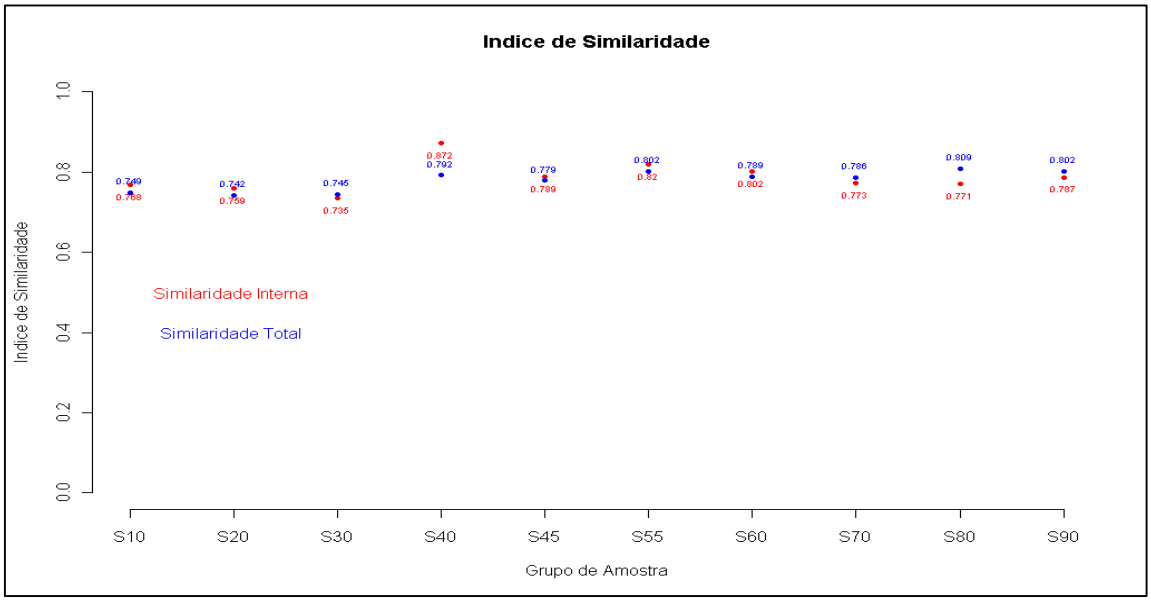


Figura 18 - Similaridade Comparativa

Como podemos notar não existe uma oscilação muito grande entre os índices de similaridades por amostra ou total.

5 Conclusões

5.1 *Geral*

Como foi proposto, foram trabalhados três conceitos o gráfico e dois numéricos. A idéia principal foi a utilização da complementariedade dos métodos numéricos acrescentando uma vertente nova, dentro do conceito visto, que foi a introdução do SAS como ferramenta de construção da optimização dos clusters.

Avaliação das Segmentações

- Avaliou-se as Segmentações obtidas tendo como pano de fundo a Perspectiva de Marketing, com o foco no critério de estabilidade interno das segmentações , diferentemente das medidas apresentadas no trabalho de Rebelo,Carmem,2006. Trabalho que nos inspirou na confecção desta dissertação buscando avaliar a estabilidade dos cluster, ponto que não foi desenvolvido por Rebelo, Carmem, 2006.

Metodologia Gráfica

Utilizando a técnica da projeção para reduzir a dimensão das variáveis, onde os gráficos obtidos não foram conclusivos em relação à verificação da estabilidade interna dos clusters encontrados, o que pode-se aferir é que com amostras ao nível de 80% conforme figura 20 existe alguma incidência de estabilidade entre os Clusters.

Metodologia Fonseca

Com base nos centróides calculamos a distância euclidiana por amostra e contra o Total, como já vimos nas Figuras 22 e 23, a avaliação que temos é que quanto menor é o % de amostragem, mais distantes se encontram os centróides, assim, analogamente nos reforça a idéia de que o CCC se "ajusta" para optimizar a quantidade de partições por amostra buscando uma homegeneidade que podemos refletir como uma estabilidade interna de cada segmento/amostra, assim, quanto menor o % de amostra existe a busca da melhor e mais homogênea aglutinação de registos, e quando se compara a base total

de dados, podemos interpretar como uma forma de compensação na busca da melhor conjugação de partições.

Metodologia Albatineh - Similaridade

O processo de calculo da similaridade está intimamente relacionado a estrutura ou podemos chamar de perfil dos segmentos, este perfil é calculado com base na semelhança de pares de registos comparados entre amostras e ou o Total. Como vemos na Figura 26 o perfil das amostras internamente ou contra o total parece não receber influência do % amostrado se de 10% ou até de 90%. Mais um ponto que nos leva a acreditar na estabilidade interna provocada pelo "ajuste" do CCC.

5.2 *Trabalho Futuro*

Com base nestes procedimentos adotados, percebemos a necessidade de pensar se em outras bases de dados o comportamento do CCC mantêm–se de forma tão equilibrada e podemos dizer até em estável. Acreditamos que como foi dito por Fonseca ET Al. A gestão da estabilidade nos segmentos pode refletir em muito no gerenciamento e redução de custos hoje tão necessário. Então um projeto futuro seria aprofundar os procedimentos adotados neste trabalho em outras bases de dados ampliando assim a metodologia aqui apresentada

Outra possibilidade seria de avaliar em campo real os clusters aqui encontrados, desta forma fazendo a ponte entre a teoria e a praxis.

6 Bibliografia

- Abonyi, J., & Feil, B. (2007). *Cluster Analysis forData Mining and System Identification.* Basel, Suiça: Springer Science.
- Albatineh, A. N., & Mihalko, D. (2006). On Similarity Indices and Correction for Chance Agreement. *Journal of Classification 23* , 301-313.
- Consulting, P. E. (s.d.). *SPSS 16 for Windows*. (W. Basic, Produtor, & SPSS Inc) Fonte: http://www.winwrap.com
- Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1990). *Introdução a Algoritmos.* MIT Press e McGraw-Hill.
- Desarbo, W. S., Grewal, R., & Scott, C. J. (2008). A Clustering Bilinear Multidimensional Scaling Methodology for Simultaneous Segmentation and Positioning Analyses. *Journal of Marketing Research , Vol. XLV*, 280-292.
- Fonseca, J. R., & Cardoso, M. G. (2007). Supermarket customres segments stability. *Journal of Targeting, Measurement and Analysis for Marketing , 15*, 210-221.
- Goeller, Susanne; Hogg, Annik; Kalafatis, Stravos P. (2002). A new research agenda for business segmentation. (Emerald, Ed.) *European Journal of Marketing , 36*, 252-271.
- Jain, A.K.; Murty, M.N.; Flynn, P.J. (1999). Data Clustering: A Review . *ACM Computing Surveys , 31*, 264-323.
- Jonhson, S. C. (1967). Hierarquical Clustering Schemes. *PSYCHOMETRYCA* , 241-254.
- Kotler, P. (2000). *Administração de Marketing, 10 Ed.* São Paulo: Prentice Hall.
- Kotler, P. (1997). *Marketing Management: analysis, planning, implementation and control.* (9a Edição ed.). New Jersey: Prentice-Hall.
- Kotler, P., & Armstrong, G. (1996). *Principles of Marketing* (7a Edição ed.). London: Prentice-Hall.
- Kumar, M., & Patel, N. R. (2007). Clustering data with measurement errors. *Science Direct - Computational Statistics & Data Analysis* (51), 6084-6101.
- Liddle, A. R. (s.d.). *Information criteria for astrophysical model selection.* Fonte: http://xxx.adelaide.edu.au/PS_cache/astrpph/pdf/0701/0701113v2.pdf
- MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). Berkeley: University of California Press.
- Mafé, C. R., & Navarré, C. L. *Determination of Number of Clusters in K-Means Clustering and Online Purchase Intention.* University of Valencia, Departament of Finance, Departament of Marketing, Valência.

- Maroco, J. (2007). *Análise Estatística com utilização do SPSS.* Lisboa: Edições Sílabo Lda.
- Ray, S., & Turi, R. H. *Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation.* Monash University, School of Computer Science and Software Engineering, Australia.
- Rebelo, C., Brito, P. Q., Soares, C., Jorge, A., & Brandão, R. (2007). *Quantitative Evaluation of Clusterings for Marketing Applications: a Web Portal Case Study.* LIAAD/INESC Porto, Faculdade de Economia, Porto.
- Rebelo, M. C. (2006). *Segmentação do Comportamento Online Utilizando Clickstream Data.* Tese de Mestrado em Ciências Empresariais, Universidade do Porto, Faculdade de Economia, Porto.
- SAS, I. (s.d.). CCC - Cubic Clustering Criterion. *SAS OnlineDoc®, Version 8* , pp. 1-39.
- Steinbach, M., Karypis, G., & Kumar, V. *A Comparison of Document Clustering Tecniques.* Tecnical Report, University of Minnesota, Departament of Computer Science and Engineering.
- Stum, A. (1982). Masther Thesis. University of College of London.
- *www.neural-forecasting.com/lvq_neural_nets.htm*. (s.d.).
- Xu, R., & Wunsch II, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks , 16*, 645-678.

7 Informações Extras / Apêndice

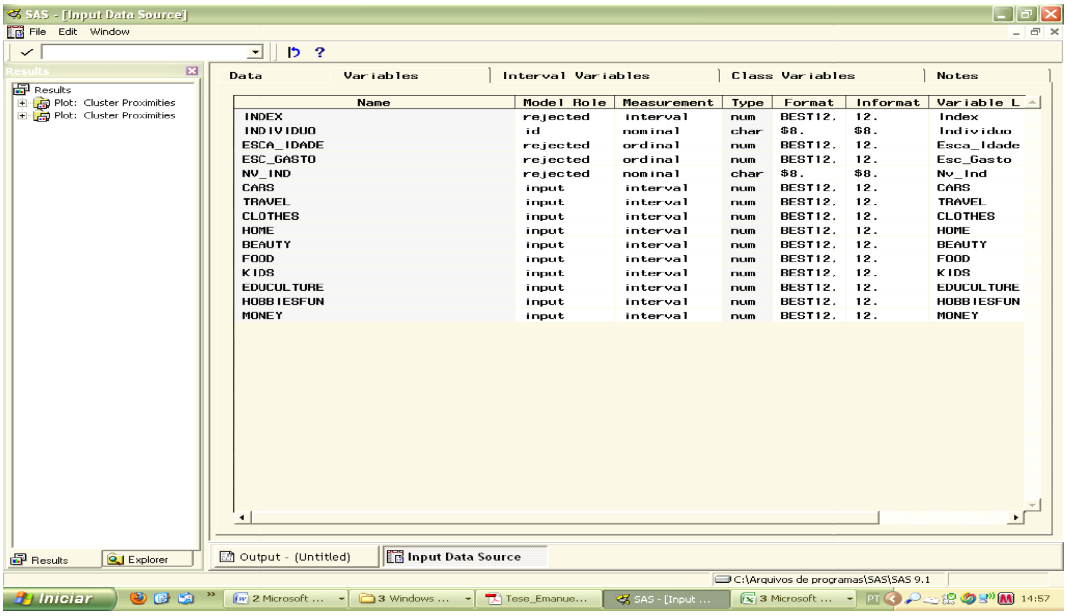
7.1 *Sementes*

Relação de sementes introduzidas no SAS e necessárias para cada % de amostra a ser retirado, exemplificando,para a amostra S1, que retiramos informação de 10%, 20%, 30%, 40%, 45%, 55%, 60%,:70%, 80% e 90% introduzimos a semente 12345, e assim por diante para cada amostra. Abaixo as devidas sementes.

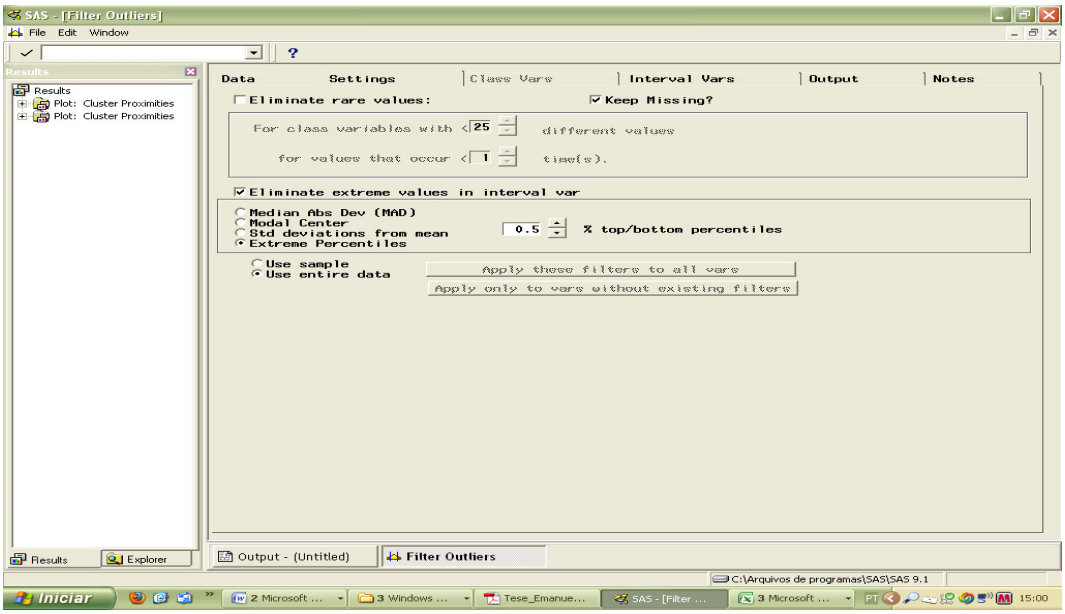
Amostra	Semente Geradora
S1	12.345
S2	753
S3	3.024
S4	1.513
S5	2.925
S6	1.433
S7	6.746
S8	9.334
S9	3.874
S10	6.881

7.2 *Telas SAS*

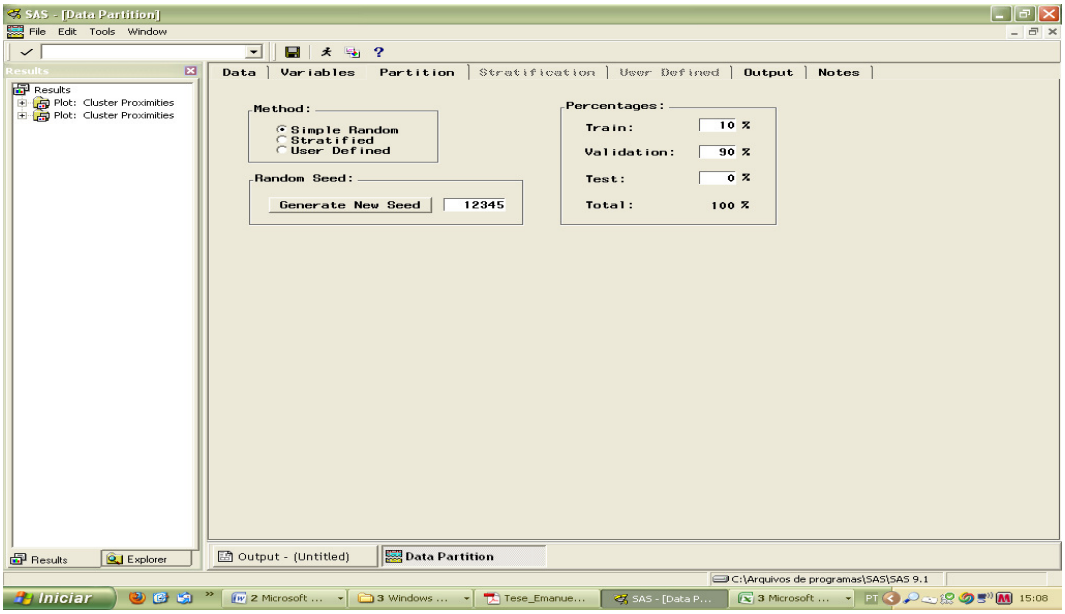
Tela de introdução das variáveis



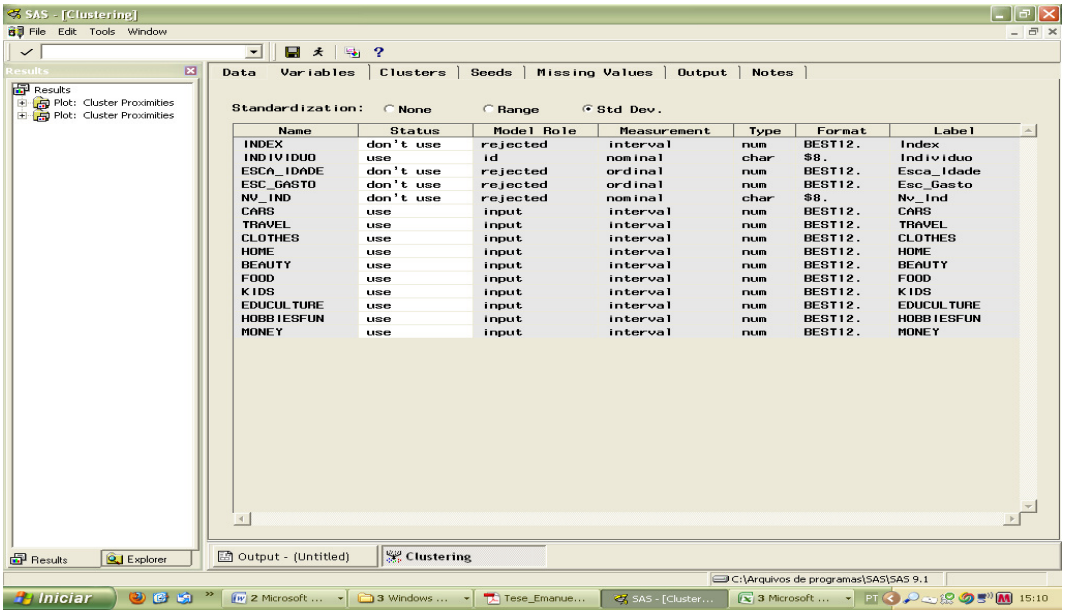
Tela de corte de Outilers



Tela de amostra aleatória simples

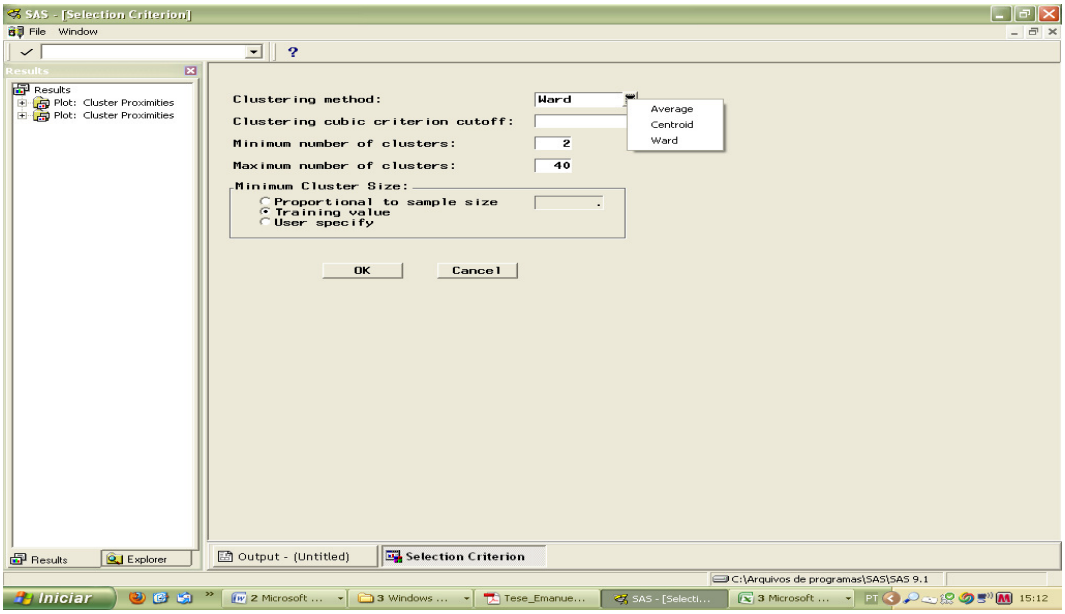


Tela de estandartização

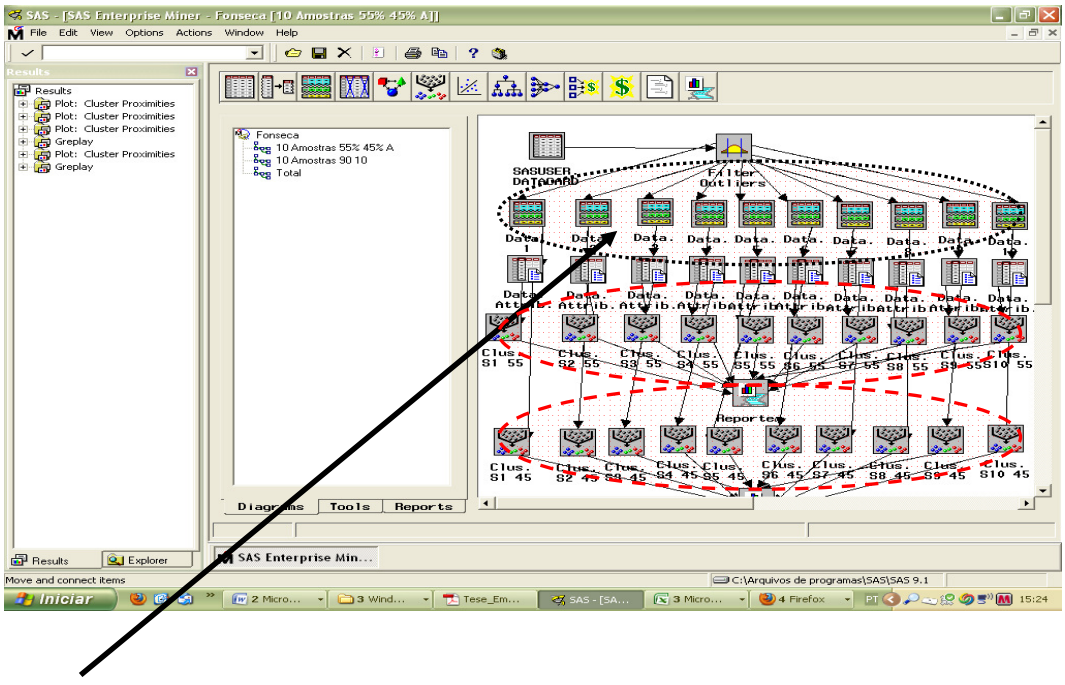


Name	Status	Model Role	Measurement	Type	Format	Label
INDEX	don't use	rejected	interval	num	BEST12.	Index
INDIVIDUO	use	id	nominal	char	$8.	Individuo
ESCA_IDADE	don't use	rejected	ordinal	num	BEST12.	Esca_Idade
ESC_GASTO	don't use	rejected	ordinal	num	BEST12.	Esc_Gasto
NV_IND	don't use	rejected	nominal	char	$8.	Nv_Ind
CARS	use	input	interval	num	BEST12.	CARS
TRAVEL	use	input	interval	num	BEST12.	TRAVEL
CLOTHES	use	input	interval	num	BEST12.	CLOTHES
HOME	use	input	interval	num	BEST12.	HOME
BEAUTY	use	input	interval	num	BEST12.	BEAUTY
FOOD	use	input	interval	num	BEST12.	FOOD
KIDS	use	input	interval	num	BEST12.	KIDS
EDUCULTURE	use	input	interval	num	BEST12.	EDUCULTURE
HOBBIESFUN	use	input	interval	num	BEST12.	HOBBIESFUN
MONEY	use	input	interval	num	BEST12.	MONEY

Tela de escolha de método



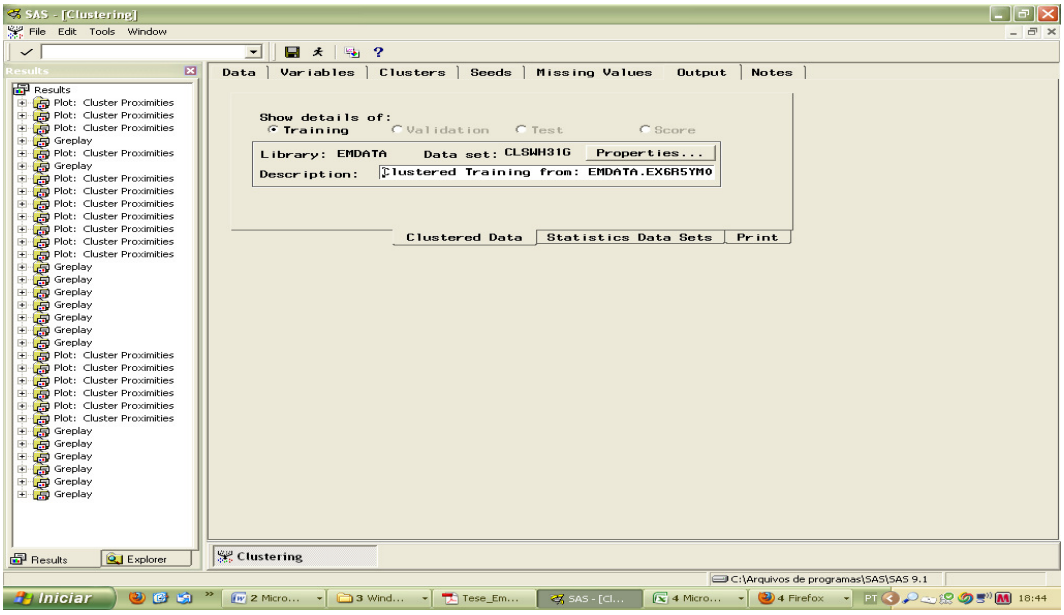
Tela com os % de cada extração para amostras de 10% e 90 %.

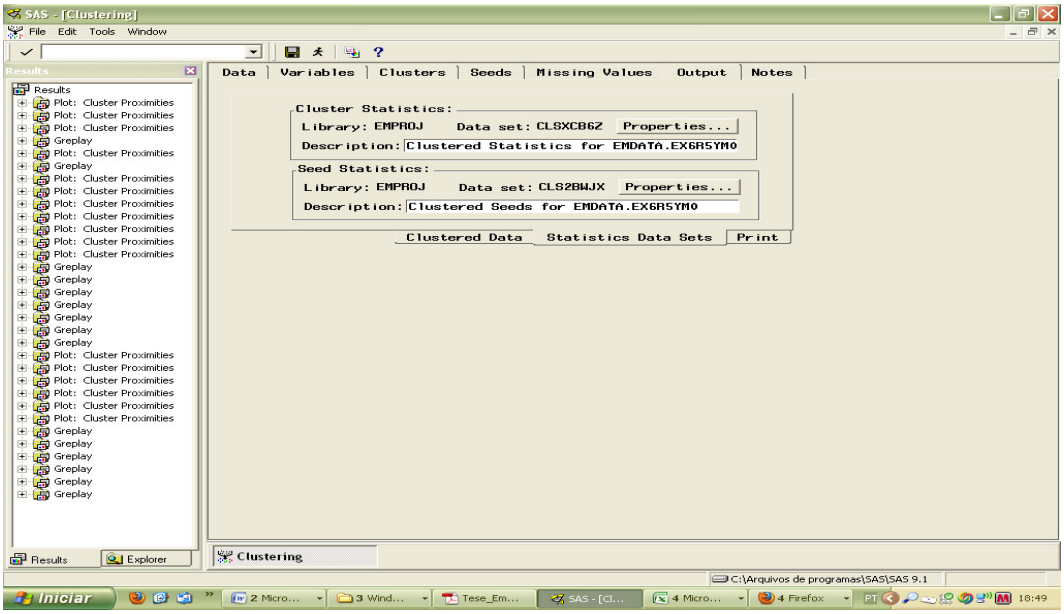


Amostras S1 a S10, utilizamos o Treino para um % e a Validação para o complemento, desta forma utilizamos todos os dados e como a aleatorização é com base

no Teste, fazer a aleatorização para 10% é diferente de utilizar o complemento da aleatorização de 90%.

Para se tratar os dados, utilizamos a exportação que o SAS fornece que é utilizando os dados que são inseridos na biblioteca EMDATA.XXXXXXXX e as estatística dos Clusters e dos Centroides na biblioteca EMPROJ.XXXXXXX





Abaixo o mapa de localização das bibliotecas que utilizamos.

S55 - Parte 1			
	Data	Cluster	Seed
S1	CLSWH31G	CLSXCB6Z	CLS2BWJX
S2	CLUSYS5U	CLSGALC6	CLSEH7LL
S3	CLUSULES	CLS2K541	CLSY19G9
S4	CLUSPY95	CLSULHXN	CLSVGTPC
S5	CLUS3GFM	CLST5XOG	CLS1F0HT
S6	CLUS63RF	CLSOGY8A	CLSWHX40
S7	CLUSGE3S	CLSH15R3	CLS21P3R
S8	CLUS5CIZ	CLSQBZ9X	CLSG4YGB
S9	CLUSJF30	CLSTZT5S	CLSBVUZJ
S10	CLUSYPTX	CLS13P2N	CLSIUOT2

S45 - Parte 2			
	Data	Cluster	Seed
S1	CLSUO3CH	CLS2VEES	CLS14LTX
S2	CLUS2W13	CLSXVIAJ	CLS0VHPX
S3	CLUS022A	CLSMA5A9	CLSJ6404
S4	CLUSIYEU	CLS1KA5S	CLSVTX0U
S5	CLUSY7H0	CLSSLXWN	CLSS41QA
S6	CLUSU2TT	CLSDBE1U	CLSVQNVP
S7	CLUS0DJY	CLSZBNB9	CLSTUJRL
S8	CLUSLCZL	CLS8280E	CLSSC0T9
S9	CLUSEB0S	CLS7HSTS	CLSL0N4O
S10	CLUSANEK	CLSP4JRD	CLSHH1NV

7.3 *Tabela das Distâncias*

Distancia Calculada pela metodologia da Distancia Euclidiana na multiplicação de matrizes de centróides.

Distancia Média Dentro das Amostras	
Rótulos de Linha	**Média de Distancia**
10	0,9576
20	0,7925
30	0,4117
40	0,5143
45	0,4526
55	0,5143
60	0,3941
70	0,4105
80	0,4438
90	0,3036

A tabela com as distancias pareadas encontradas se encontram no ficheiro: Tabela Distancia Euclidiana.pdf