USO DA TÉCNICA *TWO STEP CLUSTER* PARA SEGMENTAÇÃO DE FUNCIONÁRIOS DE UMA EMPRESA NO RIO DE JANEIRO SEGUNDO CLIMA ORGANIZACIONAL: UM ESTUDO DE CASO

Giovani Glaucio Oliveira Costa[1]

Resumo Este artigo apresenta e exploram as potencialidades de utilização de um método original e inovador de análise de dados-Técnica Two Step Cluster, particularmente apropriada à abordagem estrutural de múltiplos indicadores. Trabalha com conjuntos de dados extremamente grandes usando um algoritmo estruturado em duas etapas. Este algoritmo inova ao poder lidar com variáveis qualitativas ou quantitativas simultaneamente. No primeiro passo do procedimento, os objetos são agrupados em vários pequenos subgrupos. No segundo passo, é realizado um reagrupamento dos subgrupos, criados na etapa de pré-agrupamento, gerando os clusters propriamente ditos. Caso o número desejado de agrupamentos seja desconhecido, a análise Two Step Cluster pode automaticamente encontrar o número de clusters apropriado. Através do uso da Two Step Cluster, pode-se agrupar os dados de modo que os objetos dentro de um grupo sejam similares. Este artigo tem o objetivo de apresentar os fundamentos, os conceitos, o método, as fases de análise e a utilização da Técnica Two Step Cluster com SPSS, uma ferramenta ainda pouco conhecida e aplicada em reconhecimento de padrões, na linha de análise de segmentação de dados, exemplificando o uso e a interpretação dos resultados através de um estudo de caso: Uso da Técnica Two Step Cluster para Segmentação de Funcionários de uma Empresa no Rio de Janeiro Segundo Clima Organizacional: Um Estudo de Caso.

Palavras-Chaves: funcionários de uma empresa, segmentação, two step cluster, clima organizacional.

Abstract This article introduces and explores the potential for use of a unique and innovative method of data analysis-Two Step Cluster Technique, particularly well suited to the structural approach of multiple indicators. Working with extremely large data sets using a structured algorithm in two stages. This algorithm innovates to cope with qualitative or quantitative variables simultaneously. In the first step of the procedure, you pre-cluster the records into many small sub-clusters. In the second step, a regrouping of the subgroups authored in pre-cluster stage, generating the clusters themselves. If the desired number of clusters is unknown, the Two Step Cluster analysis can automatically find the appropriate number of clusters. Through the use of Two Step Cluster, you can group the data so that objects within a group are similar. This article aims to present the fundamentals, the concepts, the method, the analysis and the use of the technique Two Step Cluster with SPSS, a tool still little known and applied in pattern recognition, in the analysis line of slicers, exemplifying the use and interpretation of results through a case study: Use the Two Step Cluster Technique for targeting employees of a company in Rio de Janeiro the second Organizational Climate: A case study.

Keywords: air conditioned, threading, two step cluster, effectiveness.

[1]Universidade Federal Rural do Rio de Janeiro - giovaniglaucio@hotmail.com

1-Introdução

As decisões de marketing que têm por objetivo definir os melhores planos estratégicos para abordar o mercado, escolher a melhor campanha publicitária, selecionar o segmento e o tipo de produto a oferecer, têm de resultar de uma análise técnica e profissional do sistema de informação ou dos dados disponíveis. As técnicas estatísticas multivariadas conjuntamente com as diferentes soluções de software existentes hoje tornam esta tarefa em algo operacionalmente viável. Necessita-se apenas de optar pela técnica mais adequada ao objetivo do problema, selecionar uma amostra adequada para obter resultados confiáveis e posteriormente validar os resultados com outras técnicas de análise alternativas.

Muitos empresários, gestores e analistas de informações ainda encaram com reservas as pesquisas quantitativas como forma de analisar os resultados obtidos com as suas ações e opções. Receiam a confrontação entre o que desejavam e os resultados alcançados pelas empresas. Não encaram com confiança a análise de dados, por receio de não gostarem dos resultados ou por não dominarem muitas das novas técnicas de análise de dados.

Assim, desvalorizam-na embora cada vez mais esta possa ser encarada como uma base fundamental para a tomada de decisões. A utilização das novas técnicas é fundamental para comprovar que foram alcançados os resultados pretendidos com as ações e estratégias delineadas pelas empresas.

O desenvolvimento de software e a multiplicidade de soluções que existem hoje tornam cada vez mais fáceis a análise de problemas, cada vez mais complexos, se for considerado o número crescente de informação, de dados e de variáveis que envolvem cada questão organizacional e de mercado (J.D BANFIELD ET AL., 1993). Quando se trata de analisar modelos com um número cada vez mais elevado de variáveis esta tarefa tem vindo a ser facilitada. Para empresários, estudantes, etc., a informatização facilitou a análise de resultados das empresas, dos efeitos da publicidade ou ações de publicidade, nas vendas, na empatia com marcas, na imagem de instituições e no teste de eficiência de produtos e serviços postos à disposição do consumidor final (J. D BANFIELD ET AL., 1993).

Avaliar os produtos, estratificá-los segundo desempenhos diferenciados; conhecer os consumidores, as suas características, quais os mais rentáveis, que padrões de comportamento assumem e que preferências apresentam, é fundamental para melhor as empresas se adaptarem a exigências do mundo altamente competitivo. A tecnologia de hoje permite se recorrer a muitas ferramentas e técnicas para a análise de dados e facilita a interpretação de sinais de desempenho de produtos e serviços e de comportamento, preferências, de hábitos, de desejos, etc., que os consumidores apresentam, não inteligíveis numa análise direta dos valores de negócio obtido.

A segmentação de consumidores, de empresas, de marcas ou produtos permite-nos compreender melhor o mercado, quer ao nível de comportamentos e desempenho de produtos e serviços, quer ao nível de organização ou distribuição. Segmentar significa encontrar agrupamentos de indivíduos, objetos, etc., que partilhem, associem, ou seja, entendidos como tendo algumas características comuns. Com a segmentação pretende-se encontrar diferentes grupos com características homogêneas. Definir agrupamentos tem por finalidade eliminar ou diminuir a tomadas de decisão gerais aonde se aplicariam as individualizadas e personalizadas, no sentido das empresas adaptarem suas estratégias de desenvolvimento de marketing e de produto para cada subperfil evidenciado, aumentado as vendas, fidelizando marca e se destacando no mercado. Pretende-se conseguir a definição de metas, objetivos, políticas de ação ou publicidade e de desempenho adequadas a esses produtos, marcas, objetos ou consumidores.

Devem-se identificar características próprias para cada segmento: demográficas, psicográficas ou de comportamento entre consumidores, imagem, características associadas ao produto ou serviço, ao armazenamento, a eficácia, eficiência, à embalagem, à utilização dos produtos, ao hábito de compra. Valor, notoriedade, imagem, atributos de marca. Apesar de os resultados não constituírem uma clara premonição de comportamentos, uma segmentação simples pode diferenciar um conjunto de indivíduos, produtos ou objetos em grupos básicos que permitam construir parâmetros das estratégias a implementar para conseguir o tratamento, comportamentos mais adequados a cada grupo. Estes grupos podem depois, através de instrumentos mais sofisticados serem novamente desagregados ou segmentados para encontrar, identificar mais particularidades que permitam uma melhor e mais correta compreensão do seu comportamento ou necessidades.

O número de instrumentos e metodologias tem aumentado em número significativo e vem responder de forma cada vez mais eficiente aos interesses dos empresários. A análise passa então por determinar o alvo ou interesse de investigadores e empresários, o quê, quando, onde e como e depois escolher a melhor metodologia para efetuar as medidas.

Esta análise passa então primeiro pelo "Data Mining", ou seja, pela exploração dos dados. Este processo usa uma variedade de instrumentos ou ferramentas estatísticas para trabalhar os dados e encontrar padrões e relações que permitam estabelecer projeções válidas acerca do objeto do estudo (Z. HUANG, 1998). Existem várias técnicas tradicionais que os investigadores podem utilizar para atingir os seus objetivos, que vão desde a simples análise das frequências até complexas análises multivariadas (Z. HUANG, 1998).

A segmentação de consumidores de acordo com os seus comportamentos, as preferências, os hábitos, os gostos, os locais ou canais de distribuição que habitualmente visitam ou preferem; as condições que exigem aos produtos, ao desempenho, às embalagens, a imagem que associam com a marca e o valor que atribuem a esta, passam muitas vezes pela simplicidade da análise (Z. HUANG, 1998). Muitas das vezes quando o pesquisador elabora esse plano de pesquisa quantitativa, utilizando a análise estatística de dados, ele se depara com questionários onde existe uma "mistura" de variáveis nominais e contínuas numa base de dados muito grande. As técnicas tradicionais de Análise Multivariada pressupõem que todas as variáveis da base de dados sejam quantitativas e que sigam à curva normal (C. FRALEY, 1998). O analista, então, tem utilizado como alternativa o recurso analítico trivial do exame de tabelas de freqüência, tabelas de contingência com testes de independência do qui-quadrado aos dados categóricos e cálculos de estatísticas descritivas aos dados quantitativos.

A Two Step Cluster, técnica de mineração de mineração de dados recente e inovadora, pouco difundida ainda principalmente no Brasil, permite que os pesquisadores, ao utilizar questionários que envolvam perguntas com respostas nominais e/ou contínuas, saiam do estágio descritivo trivial de reconhecimento de padrões e desenvolvam técnicas multivariadas para o tratamento avançado dos múltiplos indicadores de diferentes tipos de medição em grandes bases de dados.

A Two Step Cluster vai além das análises das porcentagens das distribuições de freqüência e das tabelas de contingências e realiza análises estruturais sobre grandes bases de dados que convivem com diferentes medições

A Two Step Cluster é o instrumento analítico que privilegia as situações nas quais é necessário lidar conceitualmente e metodologicamente com objetos de estudo de configuração complexa(grandes espaços de análise com diferentes indicadores de diferentes tipos de medição). É um conjunto de critérios para analisar dados multivariados em escala nominal e/ou contínua, com o objetivo de atingir a segmentação dos registros em subgrupos homogêneos.

A Two Step Cluster busca averiguar, por via de um algoritmo escalonável, se são definidos grupos distintos. Em caso afirmativo, pretende-se averiguar como se configuram e como se posicionam os grupos uns com os outros. A Two Step Cluster é, portanto, particularmente interessante segmentação de grandes espaços de análise com configuração complexa. Com isso, a complexidade que se estabelece no espaço de análise é particionada e simplificada e configurações latentes são evidenciadas. A Two Step Cluster, então, é uma abordagem multifacetada e relacional sobre o objeto em um estudo e permite a visualização de traços de configurações num contexto de complexidade.

Aplicam-se as técnicas estruturais inovadoras da Two Step Cluster em várias áreas do conhecimento onde se tenha disponível uma grande base de dados com múltiplas variáveis em escalas diferentes e que se deseje realizar uma análise de formação de clusters para que se possa tornar mais nítida, simples e racional uma análise estrutural. Então pode ser aplicada em amplas variedades de problemas de marketing, administração, economia, contabilidade, ciências sociais, etc..

A Two Step Cluster é uma técnica nova e muito relevante no reconhecimento de padrões, na mineração de dados, quando estão envolvidos indicadores de natureza complexa, com diferentes tipos de medições, em grandes massa de informações.

Antes da disponibilização desta técnica, era complicado gerenciar ao mesmo tempo variáveis qualitativas e quantitativas simultaneamente, principalmente em segmentação de dados. Com o advento da informatização da técnica, disponível no SPSS, já é possível realizar análise de clusters de grandes bases de dados no software, levando em consideração simultaneamente múltiplas variáveis, de diferentes escalas de medição (L. KAUFMAN ET AL., 1990).

No mercado, por exemplo, pode ser indicada a Two Step Cluster para a busca da estrutura do espaço de eficácia de produtos industriais, que sejam alimentícios, eletrodomésticos, perfumaria, etc. Nesta área, a técnica pode "desenhar" as configurações do espaço em função de níveis nominais e contínuos de qualidade industrial de marcas e produtos entregues ao consumidor final.

Ao reduzir o espaço de defeitos (medidos em escala nominal e contínua) da produção industrial de uma indústria, têm-se a informação estratégica de "subperfis" de defeitos na produção e associados a estes "subperfis", deve haver um conjunto de lotes produzidos. Esta informação direciona as tomadas de decisão de fabricantes para o enfrentamento do problema da segmentação em subperfis dos defeitos gerados em lotes de produção.

Assim, Two Step Cluster pode ser utilizada em áreas diversas, desde as ciências sociais (economia, sociologia, psicologia social), às ciências humanas (história, psicologia), às ciências empresariais (administração, marketing, propaganda e opinião pública), às ciências exatas e da natureza(engenharias, qualidade, metrologia, meio ambiente, sustentabilidade, biologia, saúde) e entre outras, constituindo-se, portanto, numa abrangente técnica inovadora para a abordagem simultânea de múltiplos indicadores de diferentes níveis de medição e ao tratamento de grandes bases de dados.

O método Two Step Cluster, utilizado a partir do software SPSS versão 13.0, de análise de clusters é uma ferramenta exploratória projetada para revelar agrupamentos naturais dentro de uma base de dados que de outra maneira não se mostra aparente.

Este artigo tem o objetivo de apresentar os fundamentos, os conceitos, o método, as fases e a utilização da Two Step Cluster com SPSS, exemplificando a aplicação e a interpretação dos resultados através de um estudo de caso onde o objetivo se fundamenta na "Data Mining", isto é, na exploração de dados para otimização de estratégias de tomadas de decisão: Uso da Técnica Two Step Cluster para Segmentação de Funcionários de uma Empresa no Rio de Janeiro Segundo Clima Organizacional: Um Estudo de Caso

Resumindo, Two Step Cluster é uma técnica multivariada que tem um objetivo bem definido: o reconhecimento de padrões de subperfis coexistentes num espaço de grande complexidade.

2-Aplicações da Two Step Cluster

Quando se opta pela Two Step Cluster, o analista trabalha com grandes bases de dados, com muitas variáveis mistas coexistentes numa mesma base de dados, o que se configura num problema de configuração complexa e de multidimensionalidade. Esta constitui a principal motivação para o uso da Two Step Cluster em bases de dados criadas no SPSS, em várias áreas do conhecimento.

Na "Data Mining", a motivação para a opção pela técnica Two Step Cluster se justifica quando no espaço de partida surge a necessidade de se conhecer a estrutura que sustenta um espaço de análise de configuração complexa (Z. HUANG, 1998).

A inserção da técnica Two Step Cluster nas técnicas de segmentação é porque se sustenta na lógica de que a proximidade de certo número de registros(de diferentes tipos de variáveis) induz à presença de elementos que partilham tendencionalmente as mesmas características, isto é, têm o mesmo perfil(Z. HUANG, 1998).

Os diferentes núcleos de homogeneidade se associam a grupos de "cases" com perfis distintos, mas que coexistem com maior ou menor proximidade o mesmo espaço de partida. Quando estes núcleos dizem respeito à perfis de qualidade de produção de bens e serviços, se produzem informações estratégicas para o homem de negócios.

3-Revisão de Literatura

3.1.Definição da Técnica *Two Step Cluster*

A Análie de Clusters é uma técnica exploratória de Análise Multivariada que permite agrupar indivíduos ou variáveis em grupos homogêneos ou compactos, relativamente a uma ou mais características comuns (C. FRALEY ET AL., 1998). Dessa forma, cada observação pertencente a um determinado clusters é similar a todas as outras pertencnetes a esse clusters, e é diferente das observações pertencnetes a outros clusters.

O Two Step Cluster é um algoritmo de análise de cluster escalonável projetado para lidar com base de dados muito grandes (S. THEODORIDIS ET AL., 1999). Capaz de lidar com variáveis qualitativas e quantitativas ao mesmo tempo, requer apenas construção de uma matriz de input de dados no procedimento. Na primeira etapa do procedimento, a pré-cluster, os registros são segmentados em muitos pequenos grupos (S. THEODORIDIS ET AL., 1999). Na segunda e última etapa, a Clusterização, os pequenos grupos da etapa pré-cluster são reagrupados formando os subperfis finais segundo um número ideal de agrupamentos (S. THEODORIDIS ET AL., 1999). O método pode gerar os clusters segundo um número especificado ou se desconhecer a segmentação ideal, pode-se gerar automaticamente o número ideal de clusters para a análise (S. THEODORIDIS ET AL., 1999).

Os resultados obtidos de uma simulação de execução são consistentemente precisos e de desempenho eficáz (S. THEODORIDIS ET AL., 1999). A simulação mostra também que o procedimento automático de encontrar o número de clusters funciona muito bem e rápido (S. THEODORIDIS ET AL., 1999).

No Two Step Cluster, pode-se agrupar dados para que os registros dentro de um sbperfil sejam semelhantes (S. THEODORIDIS ET AL., 1999). Por exemplo, empresas de varejo e consumidor de produtos regularmente aplicam técnicas de clusterização de dados que

descreve seus clientes comprando hábitos, sexo, idade, renda, nível, etc. Essas empresas adaptam sua estratégia de desenvolvimento de marketing e de produto para cada grupo de consumidores para aumentar as vendas e fidelizar a marca (S. THEODORIDIS ET AL., 1999).

3.2.Metodologia da Two Steps de Cluster

Os métodos tradicionais de clusters são eficazes e precisos em pequenos conjuntos de dados, mas geralmente não são eficazes em conjuntos de dados muito grandes. Os métodos tradicionais terão desempenhos favoráveis em pequenos conjuntos de dados mas quando envolvem variáveis quantitativas contínuas. Estes são os presupostos básicos dos dois métodos tradicionais de clusters, o métodos hierárquicos e não hierárquicos (ZHANG et al., 1996).

O método do Two Steps Clsuters é um algoritmo de análise de cluster, escalonável, projetado para lidar com grandes conjuntos de dados (L. KAUFMAN ET AL., 1990). Ele pode manipular atributos ou variáveis contínuas e categóricas. Ele requer apenas a inserção dos dados no SPSS (L. KAUFMAN ET AL., 1990).

O método tem duas etapas (ZHANG et al., 1996):

1ª)Pré-cluster de casos (ou registros) em muitas pequenas porções;

2ª)Clusterização resultantes reogarnização do passo Pré-cluster para o número desejado de clusters. Pode selecionar automaticamente o número de clusters.

O algoritmo empregado por este procedimento tem diversas características desejáveis que o diferencie das técnicas de aglomeração tradicionais (ZHANG et al., 1996):

- **Manipulação de variáveis categóricas e contínuas**. Supõe que as variáveis sejam independentes, uma distribuição multinomial e normal pode ser assumida às variáveis categóricas e contínuas da modelagem;

- **Seleção automática de número de clusters**. Comparando os valores de um critério de escolha através de diferentes soluções de *clusters*, o procedimento pode determinar automaticamente o número ideal de clusters.

- **Escalabilidade.** Construindo uma árvore de recursos (CF) do *cluster* que resume os registros, o algoritmo ***Two Step Cluster*** permite analisar arquivos de dados grandes.

3.3.Medidas de Distância

Neste método, existem duas medidas de distância. Estas duas métricas determina como a similaridade entre dois clusters é calculada (ZHANG et al., 1996):

- ***Probabilidade de log*.** A medida de probabilidade ajusta distribuições de probabilidades às variáveis. As variáveis contínuas são ajustadas à Curva

Normal, enquanto variáveis categóricas são considerados de Distribuição Multinomial. Todas as variáveis são consideradas independentes.

- ***Euclidiano.*** A medida euclidiana é a distância da "linha reta" entre dois *clusters*. Ele pode ser usado somente quando todas as variáveis são contínuas.

3.4.Determinação do Número de *Clusters*

Esta selecção permite que se especifique como é determinado o número de clusters (ZHANG et al., 1996):

- **Determinação automática do número de clusters.** O procedimento pode determinarar automaticamente o "melhor" número de *clusters*, usando o critério especificado no grupo de critério de agrupamento. Comparando os valores de um critério de escolha, através das soluções de aglomeração diferentes, o procedimento pode automaticamente determinar o número ótimo de clustres. O método ***Two Step Cluster*** disponibiliza dois critérios de agrupamentos usados na dertminação automática do número de clusters para segmentar grandes bases de dados: o ***"Critério de Informação Bayesiano"*** (BIC) e o ***"Critério de Informação Akaike"*** (AIC) .

- **Especificação de um número fixo de *clusters*.** Permite fixar o número de *clusters* na solução. Deve-se digitar um número inteiro positivo.

3.5.Detalhamento da Técnica Two Step Cluster (Z. HUANG, 1998)

Nesta seção, se irá fazer uma retomada sintética dos conceitos do procedimento apresentado neste artigo para fundamentar um detalhamento não formal do algoritmo do mesmo.

Esta técnica multivariada tem como objetivo dividir elementos pertencentes a um mesmo grupo, de forma que os elementos dentro dele sejam similares entre si em relação a um conjunto de variáveis selecionadas, e os elementos em grupos diferentes sejam heterogêneos em relação a essas mesmas características. É uma ferramenta de análise exploratória de dados que tem a vantagem de agrupar objetos a partir de variáveis categóricas e não apenas de variáveis contínuas como, por exemplo, o método de k-means. As medidas de distância entre os grupos são estimadas a partir de medidas de similaridade calculadas por métodos de máxima-verossimilhança quando as variáveis são categóricas, para as quais se assume uma distribuição multinomial com variáveis independentes. Já as variáveis contínuas seguem uma distribuição normal. Apesar de esses pressupostos serem dificilmente encontrados em dados reais, o algoritmo do modelo encontra uma solução razoável, mesmo quando os pressupostos são quebrados.

O procedimento utilizado é feito em duas etapas. Na primeira, encontram-se vários pequenos subclusters e, na segunda, a partir desses subclusters, encontra-se uma resposta ótima para o melhor número de agrupamentos e os melhores clusters a segmentarem a base de dados, o qual tem como objetivo, conforme mencionado, manter a maior homogeneidade em cada grupo e a maior heterogeneidade entre os grupos. Para agrupar os elementos em cada cluster, no caso de variáveis categóricas, é adotada uma medida de distância por máxima-verossimilhança, uma medida de similaridade que mede a "distância" entre dois clusters por aproximação do decréscimo no log da função de máxima-verossimilhança.

O primeiro estágio de estimação é feito por aproximação e, para cada registro, o algoritmo toma a decisão, baseado na medida de distância, se este deve ser agrupado a algum cluster previamente formado ou se começa um novo. O objetivo é reduzir o tamanho da matriz que contém as distâncias entre os possíveis pares de clusters. O procedimento é implementado a partir da construção de uma árvore que contém as características do cluster (CF-Tree). Cada nó da árvore contém a média e a variância para variáveis contínuas e a frequência das variáveis categóricas. Para detalhes sobre a construção do algoritmo para CFTree, consultar Zhang et al. (1996). O segundo estágio parte dos subclusters formados no estágio anterior e cria o número de agrupamentos desejados através de método hierárquico aglomerativo de agrupamento. O número de grupos pode ser previamente fixado ou pode ser calculado a partir de dois critérios disponíveis – Critério de Informação de Akaike (AIC) e Critério Bayesiano de Schwarz (BIC). No presente estudo, o critério utilizado na definição do número de grupos foi o Critério Bayesiano de Schwarz (BIC).

Os resultados do método são de fácil interpretação. Os parâmetros do Critério Bayesiano de Schwarz(BIC), para uma dada rodada do método, podem ser disponibilizados através de output do Two Step Cluster. Num primeiro momento, os critérios BIC são calculados para cada número de clusters, a fim de encontrar o número inicial estimado de grupos. Num segundo momento, o número inicial de clusters é refinado. Se fosse analisado apenas o valor BIC, o critério seria selecionar o número de clusters com menor valor BIC. Contudo, o algoritmo do SPSS usa uma combinação do valor da taxa de mudança BIC "Ratio of BIC change" e da distância de máxima-verossimilhança "Ratio of Distances Measure". Assim, é selecionado o número de clusters que apresenta os maiores valores para as duas taxas mencionadas (Ratio of BIC change e Ratio of Distances Measure). Devemos mencionar que o algoritmo do SPSS não necessariamente precisa concordar com o critério do valor BIC. Quando esses critérios são divergentes, o algoritmo do SPSS julga se o ganho de informação proveniente de se ter um maior número de clusters compensa o aumento da complexidade do modelo.

A próxima informação se refere às estatísticas descritivas para as variáveis categóricas tabelas de frequência (Frequencies Table). As informações são apresentadas por cluster para cada uma das variáveis selecionadas.

Para as variáveis categóricas, o gráfico das porcentagens dentro dos clusters (Withinclusterpercentage plot) mostra como cada variável categórica é dividida dentro dos clusters. São apresentados assim, gráficos com a distribuição percentual de variáveis selecionadas por cluster.

A informação sobre a importância de cada variável na formação dos clusters (variablewise importance plot) é também apresentada graficamente. Fornecida a partir de uma medida de significância estatística (Chi-quadrado para variáveis categóricas e t-test para variáveis contínuas). Nesses gráficos temos no eixo x o valor Qui-quadrado e no eixo y a lista de variáveis. As barras que estão à direita do valor crítico indicam que as variáveis são importantes para diferenciar o cluster.

3.6.Estudos Numéricos e Simulações

O SPSS implementou a técnica Two Step Cluster em linguagem Java e C++. Testou o desempenho do método em conjuntos de dados simulados e os comparou com a simulação do desempenhos de outros métodos de clusterização (M MELIA ET AL., 1998). Os resultados mostraram que o Two Step Cluster, desenvolvido pelo SPSS, é diferente de qualquer método existente semelhnte hoje (M MELIA ET AL., 1998). Foi o mais eficaz e o mais eficiente na formação de grupos homogêneos e o de melhor desempenho na determinação automática de

clusters para bases de dados, inclisive as que são formadas pos grande número de registros (M MELIA ET AL., 1998).

A Two Step Cluster, então, é uma ferramenta exploratória para revelar subperfis em bases de dados onde a "olho nu" não seria aparente. O algoritmo empregado por este procedimento tem várias características desejáveis que o diferenciam das técnicas tradicionais de clusters.

3.7.Sequência de Passos no SPSS para Realização do *Two Step Cluster*:

Sequência de Passos:

1. *ANALYSE;*
2. *CLASSIFY;*
3. *TWOSTEP CLUSTERS;*
4. *Em **"CATEGORIAL VARIABLES",** inserir as variáveis qualitativas do modelo. Em **"CONTINUOS VARIABLES"**, inserir as variáveis quantitativas do modelo;*
5. *Em **"NUMBER OF CLUSTERS"**, se especificará o número de clusters a se criar. Se o analista desconhecer o número de clusters ideal a segmentar a base de dados, assinar a opção **"DETERMINE AUTOMATICALLY"**. O "default" é gerar até 15 agrupamentos.*
6. *Em **"PLOTS"**, assinalar as opções: **"WITHIN CLUSTERS PERCENTAGE CHART", "CLUSTERS PIE CHART".*** Em **"VARIABLE IMPORTANCE PLOT"**, assinalar "**RANK VARIABLES"**, em seguida marcar as alternativas **"RANK VARIABLES→BY "CLUSTER" ou "BY VARIABLE", "IMPORTNCE MEASURE→CHI-SQUARE OR t-TEST OF SIGNIFICANCE" e "CONFIDECE LEVEL".**
7. *CONTINUE;*
8. *OK.*

Efetuada esta sequência de passos, o analista terá como output uma tabela de distribuição de frequência de cada clusters, uma tabela com as médias assumidas pelas variáveis quantitativas do modelo em cada clusters(quando o estudo tiver variáveis quantitativas), distribuição de frequência das categorias de cada variável categórica do estudo em cada cluster(quando o estudo tiver variável qualitativa), gráfico de "pizza" das freqüências dos clusters, gráfico de barras para percentagens dentro dos clusters para as variáveis qualitativas, gráfico dos intervalos de confiança para as variáveis quantitativas, gráfico de significância do qui-quadrado para as variáveis qualitativas do modelo e finalmente gráfico de significância t-Student para as variáveis quantitativas.

Observações:

- Se o analista desejar que o gráfico de significância do *qui-quadrado* e o gráfico de significância t-*Student* omita as variáveis **não significantes** deverá assinalar em "CONFIDECE LEVEL" a opção "OMIT INSIGNIFICANT VARIABLES";

- Se o analista desejar obter os resultados da simulação que determinou, por um critério especificado, o número automático de *clusters*, deverá assinalar a sequência **"OUTPUT→STATISTICS→INFORMATION CRITERION(AIC OE BIC).**

4-Estudo de Caso: *Uso da Técnica **Two Step Cluster** para Segmentação de Funcionários de uma Empresa no Rio de Janeiro Segundo Clima Organizacional: Um Estudo de Caso*

4.1-Fundamentação Teórica do Caso

Conceitua-se como "Clima Organizacional" a ferramenta administrativa, integrante do Sistema da Qualidade, utilizada para medir e apurar o grau de satisfação dos colaboradores diretos de uma empresa perante determinadas variáveis. Pode ser utilizada e aplicada isoladamente ou de forma conjunta com as demais ferramentas do Sistema de Qualidade.

A direção da empresa deve determinar, com o auxílio de especialistas na área de comportamento e relacionamento social, além de técnicos da área de recursos humanos, o que acredita ser um resultado padrão ideal da satisfação de seus funcionários.

Para a determinação desse padrão ideal são considerados aspectos variados como: satisfação com salário, reconhecimento e promoção, clima entre os funcionários e o clima entre funcionários e chefia que podem ser tomados como indicadores do "Clima Organizacional".

Determinado o ponto padrão de satisfação, tido como ideal pela empresa, admite-se uma variação percentual para mais ou para menos, que representaria uma faixa de tolerância aceitável como satisfatória. Caso a pesquisa aponte como resultado final algum ponto enquadrado na faixa de tolerância, conclui-se que o "Clima Organizacional" está satisfatório.

Como a grande maioria das ferramentas da qualidade, o "Clima Organizacional" também é uma ferramenta estatística e, a exemplo das demais, procura detectar e apontar distorções nos processos administrativos, para análise e ponderações da direção das empresas que, se for o caso, adotará medidas corretivas, antecipando-se a futuros problemas. No caso do "Clima Organizacional", problemas relacionais.

Note que o padrão ideal do "Clima Organizacional" é determinado pela direção da empresa, o que nem sempre agrada e atende aos anseios de seus empregados. Algumas empresas já evoluíram nesse aspecto e determinam o ponto padrão de comum acordo com os colaboradores, que são representados, no processo, por equipes formadas por colaboradores indicados pelo quadro funcional.

Para que o processo de estabelecimento do ponto ideal do "Clima Organizacional" seja utilizado corretamente, a direção da empresa e a equipe representativa dos colaboradores devem estar conscientes, totalmente engajadas e desprovidas de quaisquer bloqueios.

Há necessidade de ambas as partes se conscientizarem de que precisam encontrar uma solução que atenda a todos os interesses, ou seja, "a chefia precisa do empregado tanto quanto o empregado precisa do patrão e de seu emprego". As equipes deverão buscar de forma racional, madura e negociada, o ponto de equilíbrio que satisfaça as partes.

Para isso, realizam-se pesquisas internas nas quais se submete, periodicamente, aos funcionários diretos, independentemente do posto, função e do nível hierárquico na empresa, questionários com opções de respostas pré-definidas para a apuração do grau de satisfação. Determina-se, então, o que seria aceitável como padrão de satisfação e analisam-se os resultados.

Portanto, o "Clima Organizacional" é mais uma ferramenta disponível para utilização pelo administrador no processo organizacional e administrativo das empresas. Essa ferramenta bem utilizada possibilita à direção das empresas detectarem fatores, procedimentos adotados e alguns problemas de relacionamento hierárquico e relacional que não são bem aceitos pelos seus colaboradores diretos, o que pode representar ponto de atrito e

estrangulamento no processo organizacional, relacional e evolutivo das empresas e afetar os negócios.

Cabe ao administrador analisar e ponderar criteriosamente a análise dos dados para verificar algum ponto que necessite ser corrigido para que o processo produtivo da empresa se mantenha isento de incompatibilidades relacionais e insatisfações pessoais.

O "Clima Organizacional" é uma ferramenta mais comumente utilizada em grandes empresas que, por seu tamanho, não permitem à direção detectar distorções setoriais e departamentais que possam estar afetando seu desempenho organizacional. Problemas relacionais não são raros. Muitas empresas já vivenciaram problemas como, por exemplo, determinado setor ou departamento que apresenta problemas de relacionamento e satisfação individual ou coletiva, afetando, assim, o resultado final da empresa.

Essa ferramenta é extremamente importante quando é utilizada de maneira correta, imparcial e isenta de paixões ou sentimentos pessoais. Quando utilizada de forma periódica, pode proporcionar à direção da empresa a tomada de medidas administrativas proativas que evitem o surgimento de problemas ou seu agravamento.

A ferramenta "Clima Organizacional" é tão importante que a maioria dos órgãos certificadores de excelência empresarial e organizacional já incorporou ao processo de certificação a exigência de seu uso no processo de qualificação das empresas.

Entretanto, grande parte das empresas que adotaram essa ferramenta para o fim de certificação de excelência, principalmente empresas estatais e órgãos públicos, utilizam a ferramenta de forma limitada, verificando-se o "Clima Organizacional" à nível geral, macro, quando na verdade podem coexistem dentro da empresa diferentes climas organizacionais, referentes a diferentes grupos de colaboradores.

Ao se tomar uma estratégia geral de ação para melhorar o clima organizacional na empresa, quando existe uma estratificação de satisfação, corre-se o risco de desenvolver planos de correção que não atendam a todos os segmentos de colaboradores latentes e a todos os indicadores de clima organizacional, mas sim a uma porção mais geral.

É pertinente, então, escolher um método de análise de dados que atende ao objetivo de se segmentar os funcionários da empresas em função de seus possíveis diferenciados climas de satisfação.

Na elaboração de um plano de pesquisa que preveja uma coleta extensa de informações, a seleção dos métodos de análise de dados é uma tarefa de importância decisiva, quaisquer que sejam os fenômenos em estudo. É essencial garantir que os instrumentos que venham a ser selecionados permitam uma análise adequada dos dados, sejam consistentes com a natureza das informações colhidas e atendam ao objetivo do estudo.

A Two Step Clusters é um instrumento analítico que privilegia as situações nas quais é necessário lidar conceitualmente e metodologicamente com objetos de configuração complexa. A Two Step Cluster consiste numa abordagem multifacetada e relacional sobre o objeto em estudo indicando traços de sua configuração complexa.

A Two Step Cluster é particularmente interessante para segmentar grandes bases de dados onde se afiguram variáveis com diferentes níveis de mensuração.

A Two Step Cluster constitui, portanto, num conjunto de critérios para analisar dados multivariados em adversidade de medição, com o objetivo de se criar grupos homogêneos dentro e heterogêneo entre. A técnica trata múltiplas variáveis categoriais e contínuas analisadas conjuntamente e desenvolve análises estruturais. É um método de análise quantitativa para variáveis qualitativas e quantitativas e tem por objetivo atingir a solução ótima na criação de clusters ou subperfis.

Esta seção objetiva realizar uma mineração de dados, um reconhecimento de padrões de funcionários de uma grande empresa no Rio de Janeiro, com vistas a sua segmentação, em função de variáveis que meçam o clima organizacional.

Foi selecionada uma amostra aleatória simples de 200 funcionários da empresa e a essa amostra foi aplicado um questionário estruturado com cinco variáveis: salário, reconhecimento e promoção, clima entre funcionários e clima entre funcionários e chefia.

Neste estudo, desde o início, tanto os funcionários, quanto a empresa, foram mantidos em sigilo para se garantir que não houvesse problemas de precisão da pesquisa e possíveis questões de conflitos futuros de "assédio moral".

Para realizar a clusterização da base de dados optou pela técnica Two Step Cluster, uma vez que essa é a técnica de análise de cluster que se potencializa quando se está diante de variáveis medidas em diferentes níveis de mensuração e com grandes bases de dados.

Portanto, a Two Step Clusters é a hipótese preliminar de recurso de análise apropriado para se atingir o objetivo deste estudo de caso e que será explorado e prestigiado no desenvolvimento das seções a seguir.

Os parágrafos a seguir tratarão do desenvolvimento das etapas do método Two Step Clusters. A rodada do Two Step Cluster foi realizada no SPSS 15.0, mas este método já está disponível a partir do SPSS 13.0.

4.2-Desenvolvimento do Método *Two Step Cluster*

1º) Preparação da Matriz de *Input*

O objetivo deste estudo é realizar uma análise estrutural dos funcionários de uma grande empresa, realizando a investigação de possíveis subclimas organizacionais.

A amostra utilizada abrangeu 200 funcionários que responderam a um questionário estruturado onde teriam que dá uma nota de 0 a 10 para seu grau de satisfação com itens como salário, reconhecimento e promoção, clima entre funcionários e clima entre funcionários e chefia. No questionário, foi coletado o cargo de cada funcionário, que é apresentado na Tabela 4.1.

Na Tabela 4.1, as variáveis são detalhadas e as categorias são codificadas. Depois de codificadas as categorias, o que se obtém é a "Matriz de Input" dos dados para o SPSS, para realização da Two Step Clusters.

Tabela 4.1- Variáveis da Pesquisa e Codificação das Categorias

Variáveis	Descrição e Codificação
Salário	Avaliação do salário numa escala de 0 a 10
Reconhecimento e Promoção	Avaliação do reconhecimento e promoção numa escala de 0 a 10
Clima entre os funcionários	Avaliação do clima entre os funcionários numa escala de 0 a 10
Clima entre os funcionários e chefia	Avaliação do clima entre os funcionários e chefia numa escala de 0 a 10
Cargo ocupado na empresa	Cargo ocupado na empresa(A= Apoio e Operacionais, B=Gerentes e Técnicos Especializados, C=Alta Administração
Variáveis	Descrição
Salário	Avaliação do salário numa escala de 0 a 10
Reconhecimento e Promoção	Avaliação do reconhecimento e promoção numa escala de 0 a 10
Clima entre os funcionários	Avaliação do clima entre os funcionários numa escala de 0 a 10
Clima entre os funcionários e chefia	Avaliação do clima entre os funcionários e chefia numa escala de 0 a 10
Cargo ocupado na empresa	Cargo ocupado na empresa(A= Apoio e Operacionais, B=Gerentes e Técnicos Especializados, C=Alta Administração

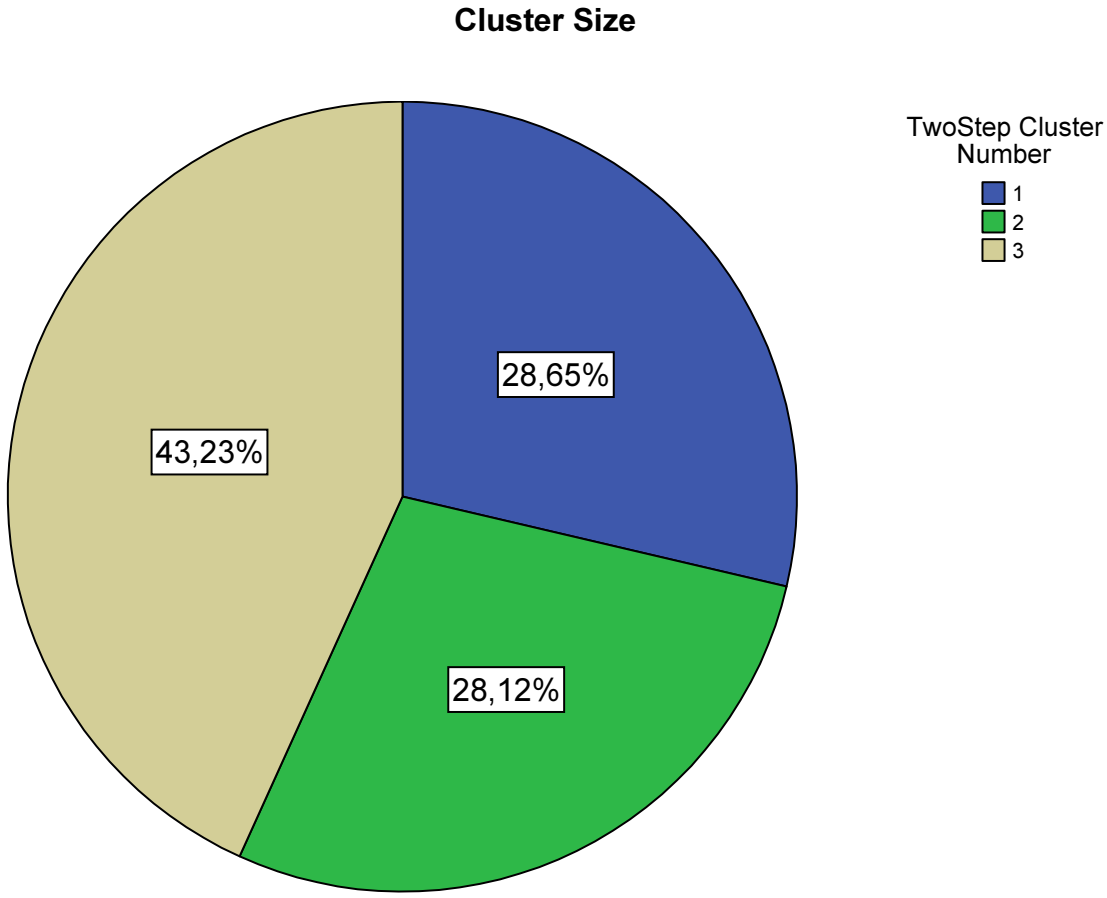
2º)Análise da Frequência de cada Cluster

Tabela 4.2 – Distribuição dos Clusters

Clusters Formados	Tamanho dos Clusters	% de Objetos Combinados	% doTotal
1	55	28.6%	27.5%
2	54	28.1%	27.0%
3	83	43.2%	41.5%
Combinados	192	100.0%	**96.0%**
Casos Excluídos	8		4.0%
Total	200		100.0%

Pelo que se pode observar da Tabela 4.2, que 96,0% funcionários da empresa foram combinados em alguns clusters. Portanto, no presente estudo, foi possível classificar em grupos homogêneos quase a totalidade dos funcionários, que estão separados em três clusters com as freqüências apresentadas na tabela referida.

Gráfico 4.1-Tamanho dos Clusters

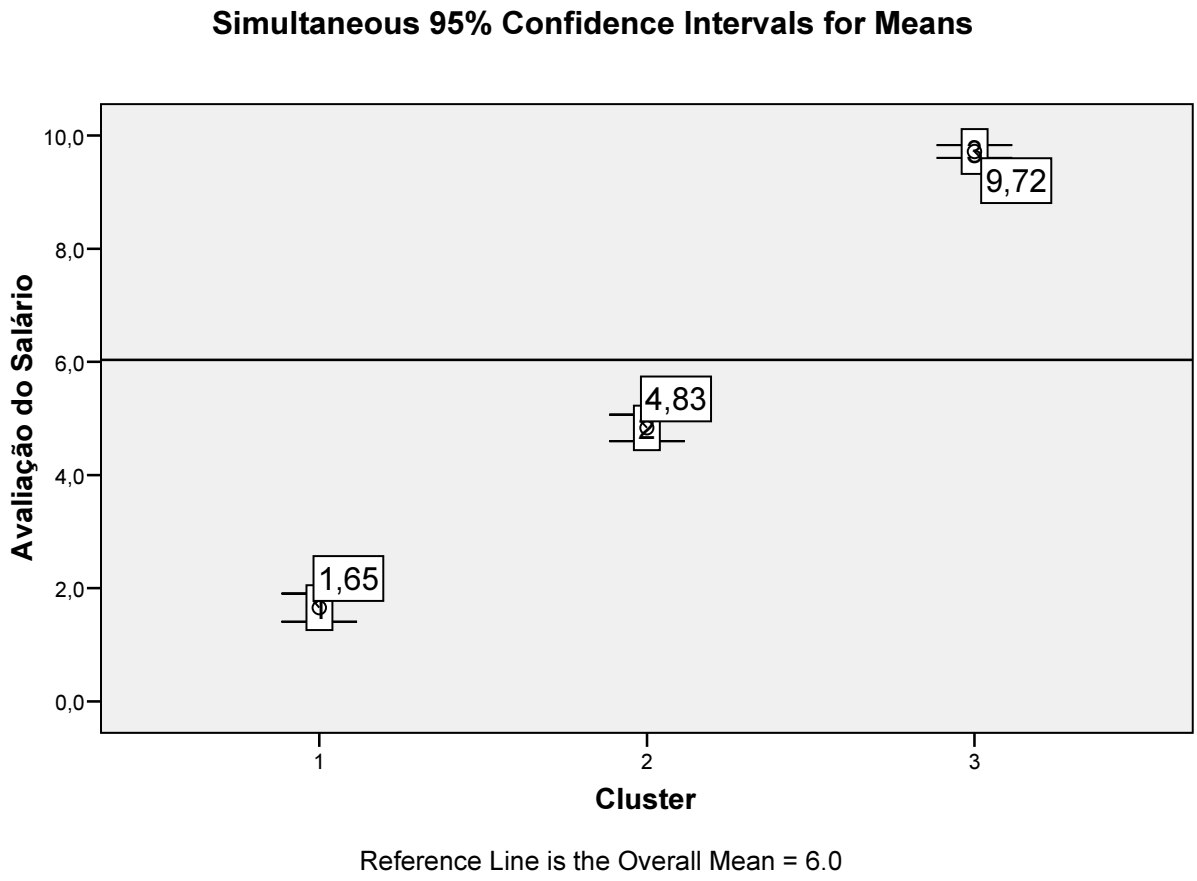


O Gráfico 4.1 revela instantaneamente que o Clusters 3 é o que congrega a maioria das funcionários (43,23%). O interessante é avaliar dentro do foco “clima organizacional”, qual o perfil destes clusters.

3º)Análise do Comportamento das Variáveis pelas suas Categorias Dentro dos Clusters

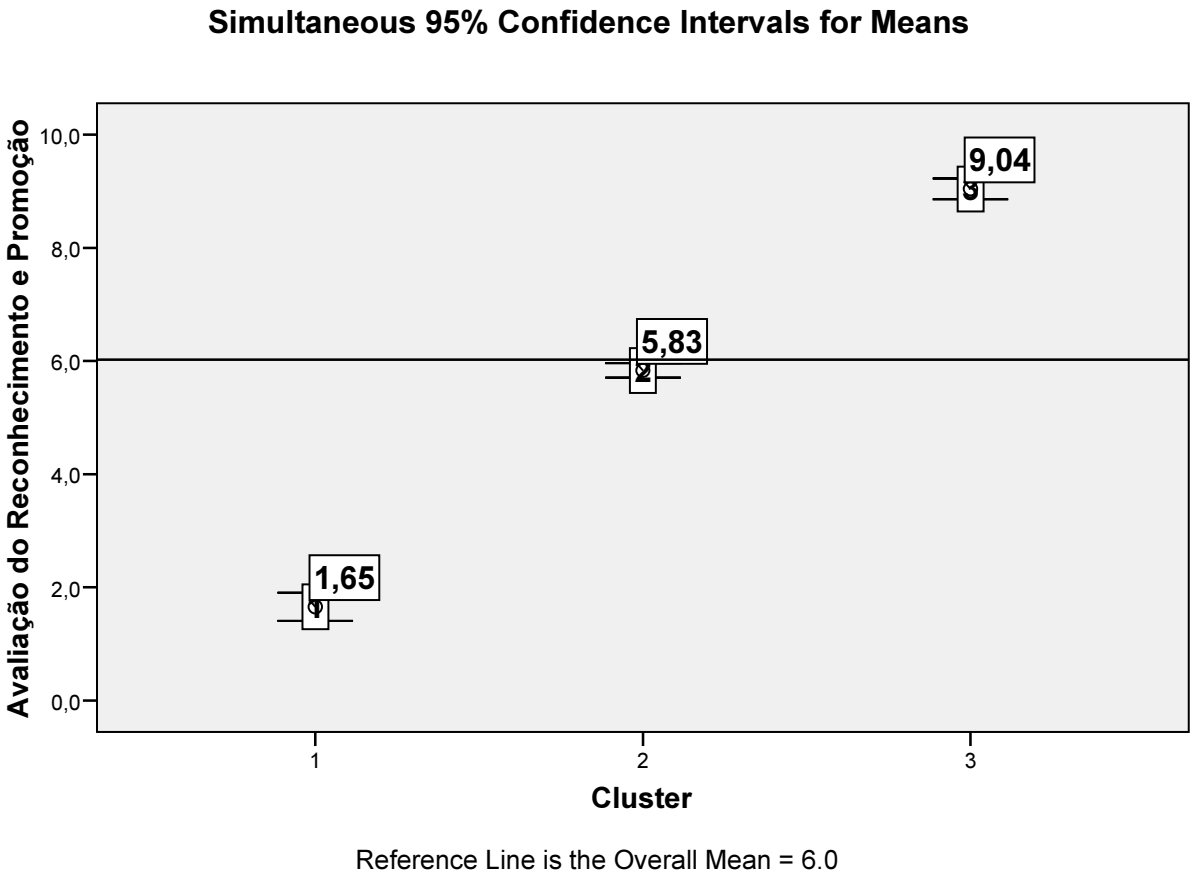
Gráfico 4.2- Percentagem das Categorias das Variáveis dentro dos Clusters

Salário



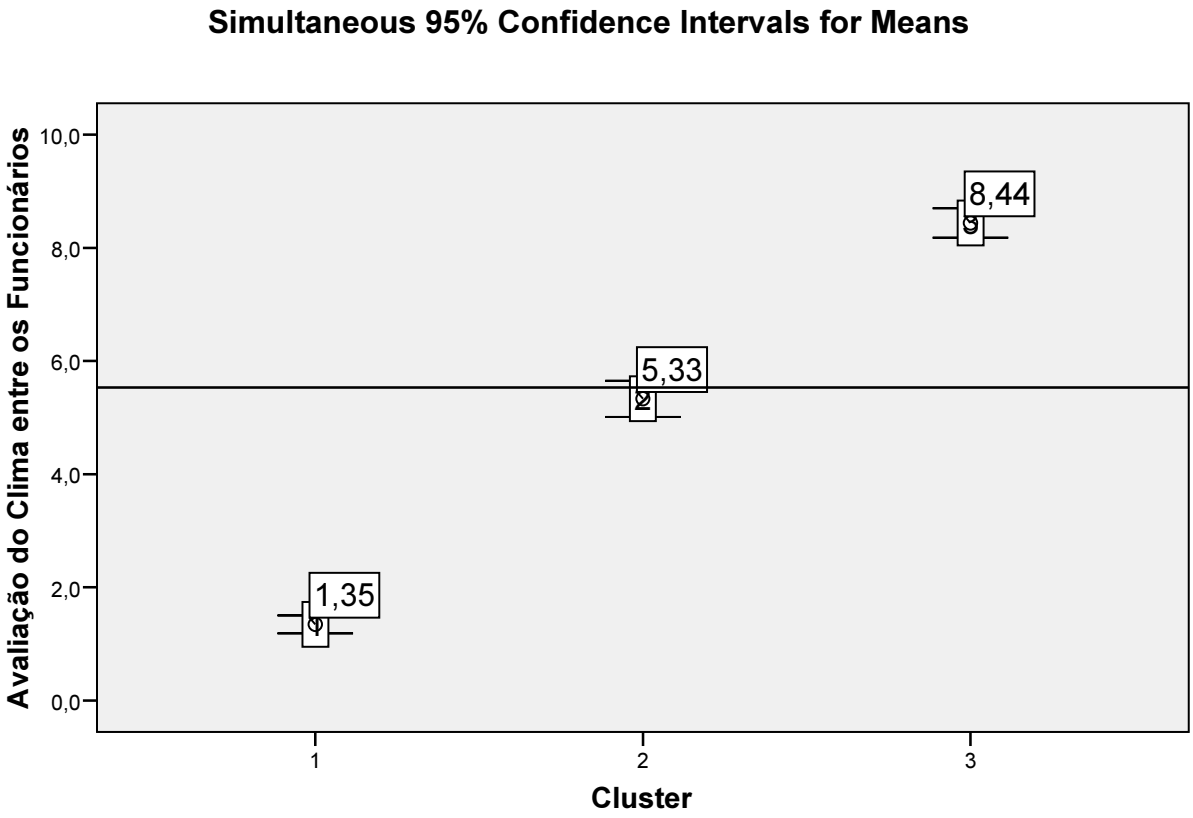
Observando o gráfico acima, quanto ao salário, o grau de satisfação é crescente do cluster 1 ao cluster 3.

Reconhecimento e Promoção



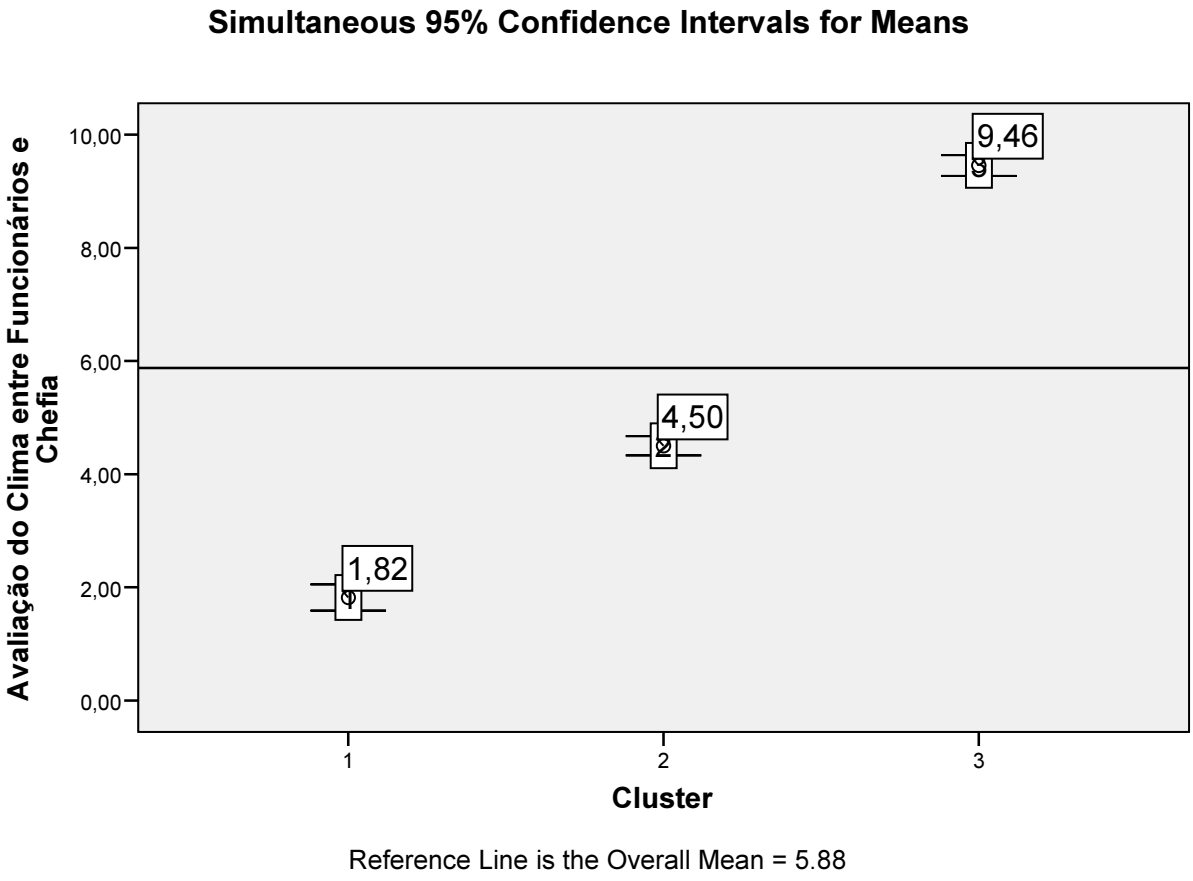
Observando o gráfico acima, quanto ao reconhecimento e promoção, o grau de satisfação é crescente do cluster 1 ao cluster 3.

Clima entre Funcionários



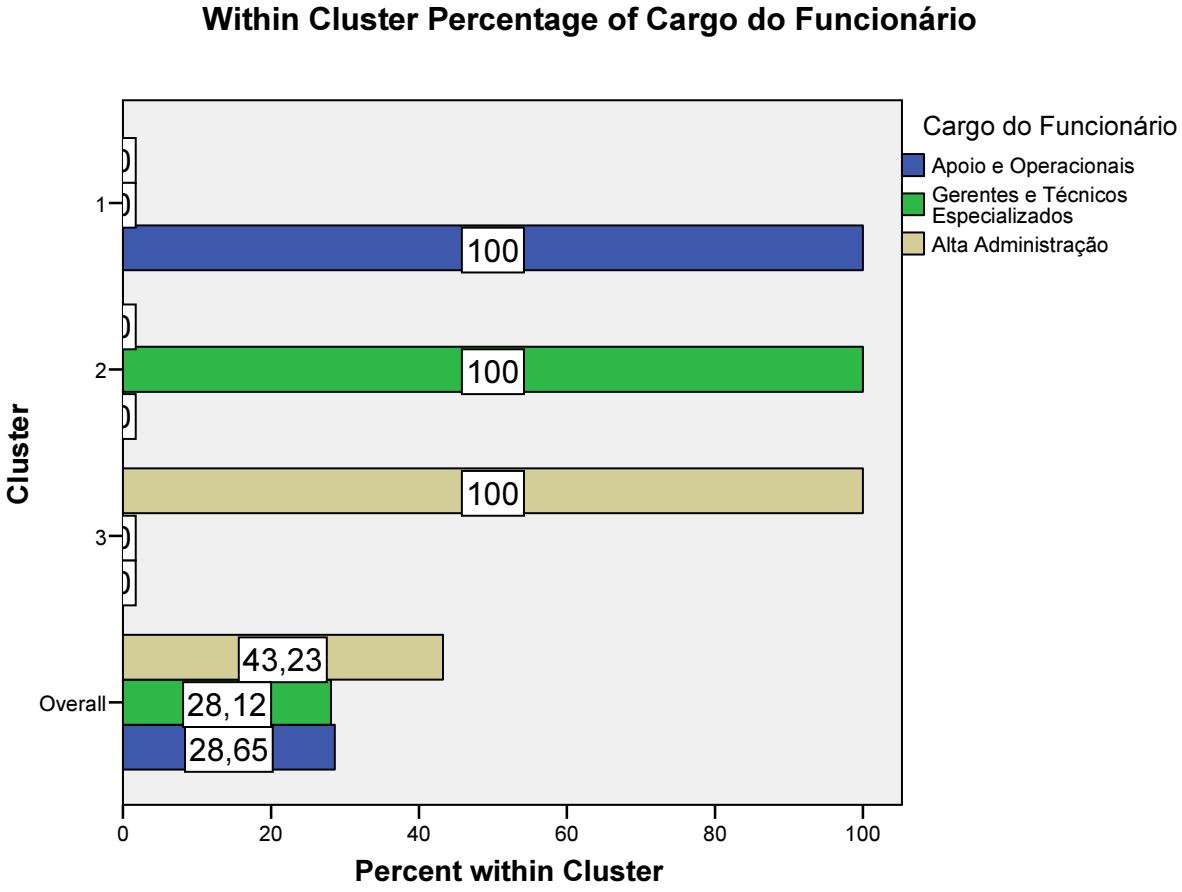
Observando o gráfico acima, quanto ao clima entre funcionários, o grau de satisfação é crescente do cluster 1 ao cluster 3.

Clima entre Funcionários e Chefia



Observando o gráfico acima, quanto ao clima entre funcionários e chefia, o grau de satisfação é crescente do cluster 1 ao cluster 3.

Cargo Ocupado na Empresa



Observando o gráfico acima, o cluster 1 é formado pelos funcionários com os cargos de apoio e promoção. O cluster 2 é formado totalmente pelos funcionários nos cargos de gerentes e técnicos especializados e finalmente o cluster 3 é constituído pelos funcionários da alta administração.

Através do Gráfico 4.2 é possível traçar o perfil de todos os clusters ou configuração, antes latente na base de dados.

O cluster 1 é formado pelos funcionários que deram avaliações mais baixas ao salário, ao reconhecimento e promoção, ao clima entre funcionários e ao climas entre funcionários e chefia. São os funcionários que ocupam o cargo de apoio e operacional.

O cluster 2 pelos funcionários que deram notas medianas ao salário, ao reconhecimento e promoção, ao clima entre funcionários e ao climas entre funcionários e chefia. São os funcionários que ocupam o cargo de gerência e técnicos administrativos.

Finalmente, as maiores notas aos quesitos de clima organizacional da empresas em estudos foram os da alta administração, cluster 3.

A análise revelou, por conseguinte, que existem evidências de subclimas no espaço investigado que se distanciam do perfil geral existente e que a visão que se deve ter do problema deve ser focada em cada configuração identificada.

Enfim, chegou-se à estruturação em perfis da base de dados, isto é, foram identificados os três subclimas que se coexistem na amostra estudada. A análise atual revelou,

então, como inicialmente se questionou neste estudo, que existe variação em torno do clima organizacional geral e que três configurações possuem perfis de satisfação com as empresas diferentes, mas que coexistem num mesmo espaço de análise. Com uma mineração, um reconhecimento de padrões, efetiva dos dados, detectou-se níveis distintos de regularidade do problema estudado.

A informação do perfil de cada cluster constituintes da base de dados sugeriria tomadas de decisão diferenciadas para cada grupo insatisfeito, para minimizar o problema de intolerável clima organizacional na empresa.

Com a análise concluída, os administradores da empresa têm a informação do clima organizacional de cada cargo da empresa e podem adaptar as suas estratégias de desenvolvimento de recursos humanos para otimizar o nível de satisfação da empresa, minimizando as insatisfação segmentárias e fazendo convergir para um único grau satisfatório de clima organizacional bom para todos os funcionários da empresa.

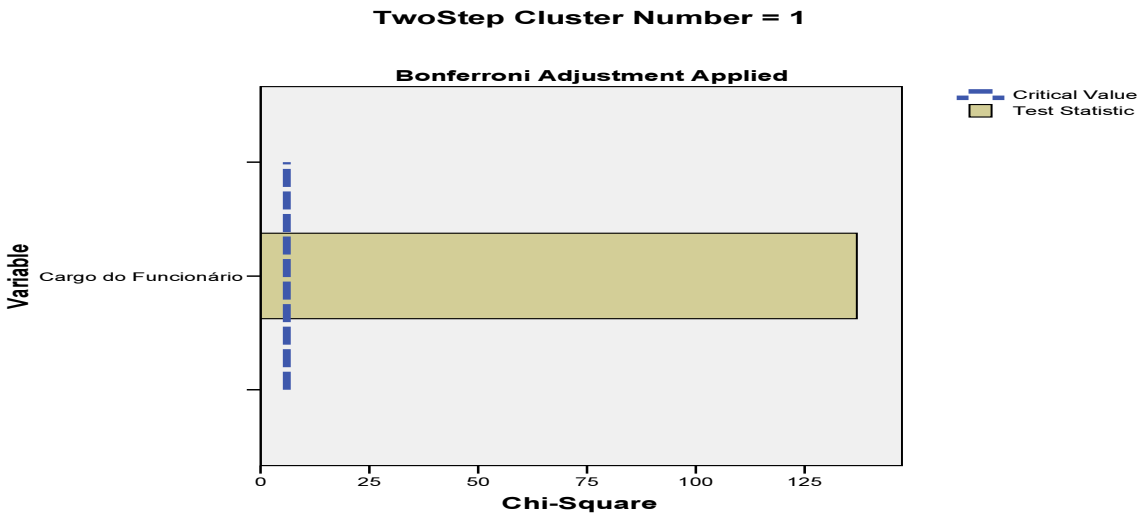
4º)Teste de Significância das Variáveis em cada Clusters

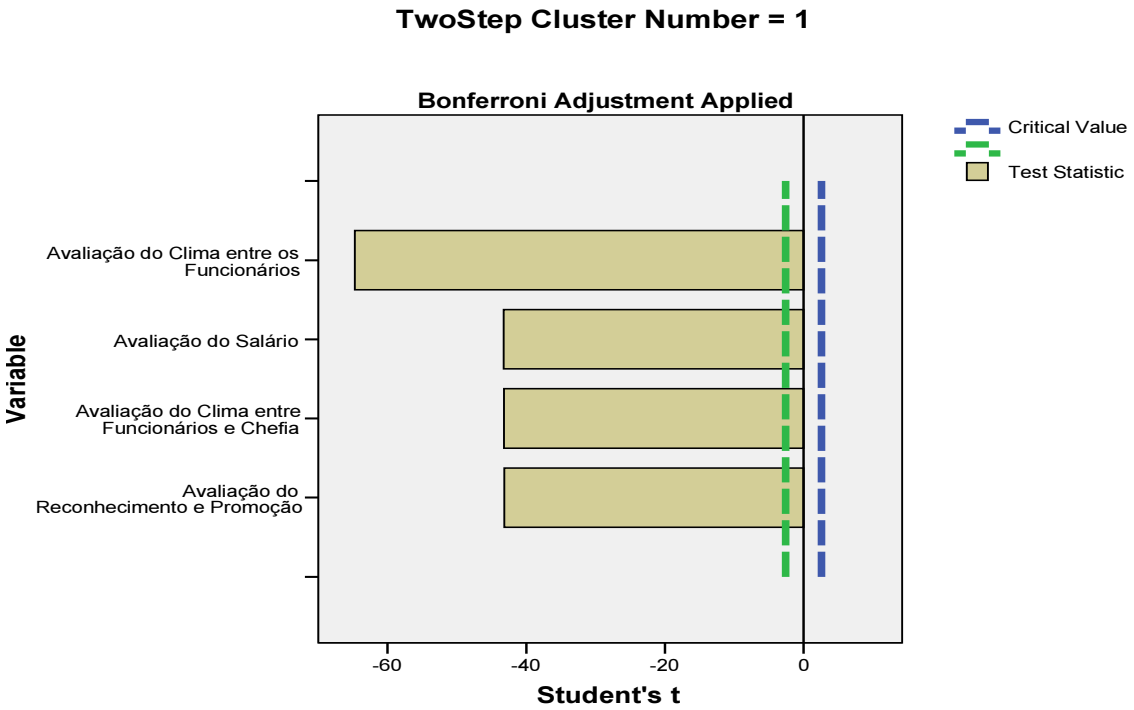
Após a análise de clusters, os dados devem ser submetidos a testes de significância estatística para avaliar quais variáveis que tiveram importância na formação dos clusters. No Gráfico 4.3, estão ordenadas as variáveis de acordo com a relevância que elas tiveram no processo de classificação dos grupos. Essa ordem foi apontada através dos resultados do teste de significância.

Gráfico 4.3- Gráfico de Importância das Variáveis nos Clusters

Variáveis Importantes no Clusters 1-Teste de Significância para o Cluster 1

No Cluster 1:



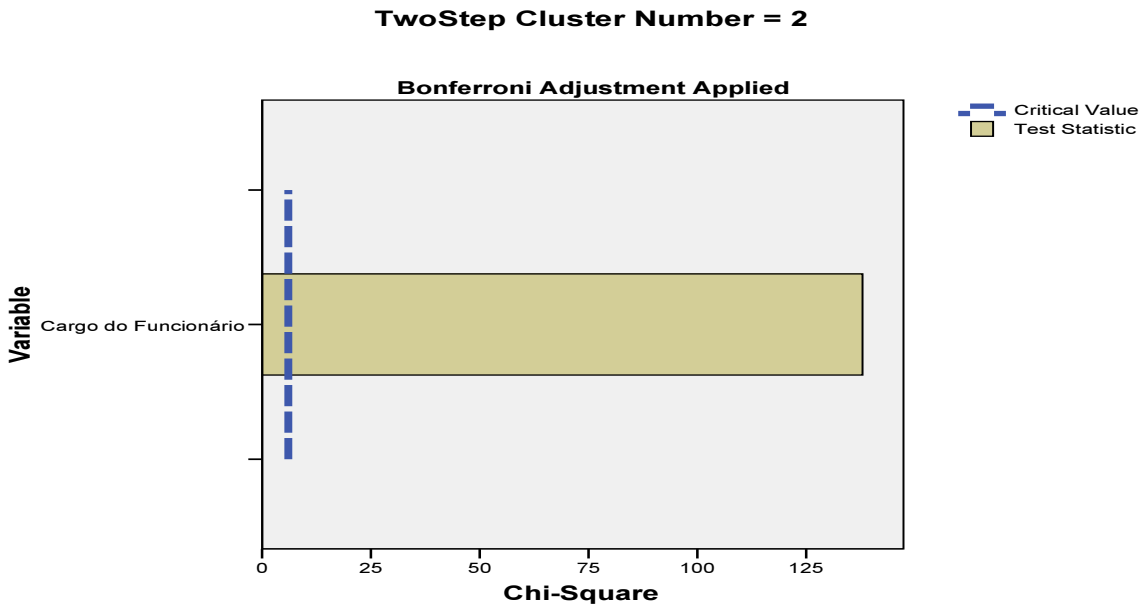


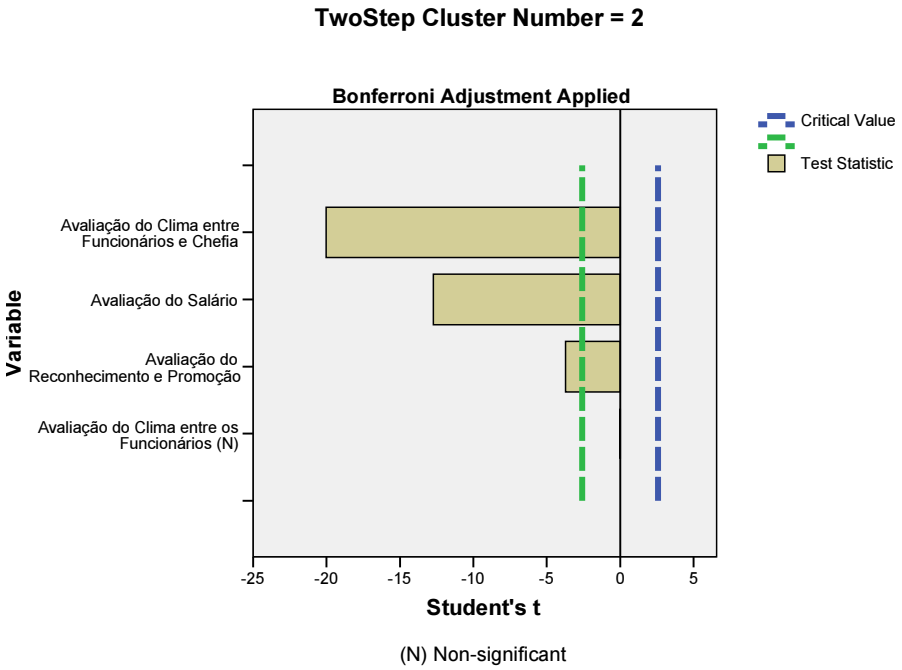
Decisão:

Todas as variáveis da análise são significantes estatisticamente para a formação do cluster 1, isto é, discrimina-o e são importantes para a sua formação.

Variáveis Importantes no Clusters 2-Teste de Significância para o Cluster 2

No Cluster 2:

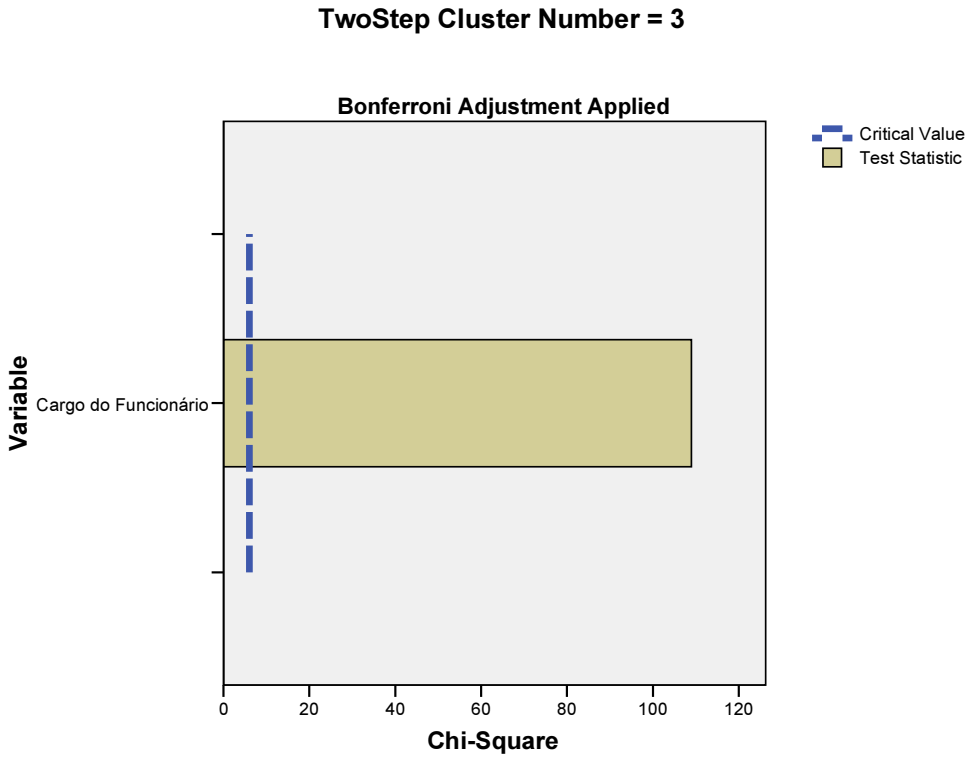


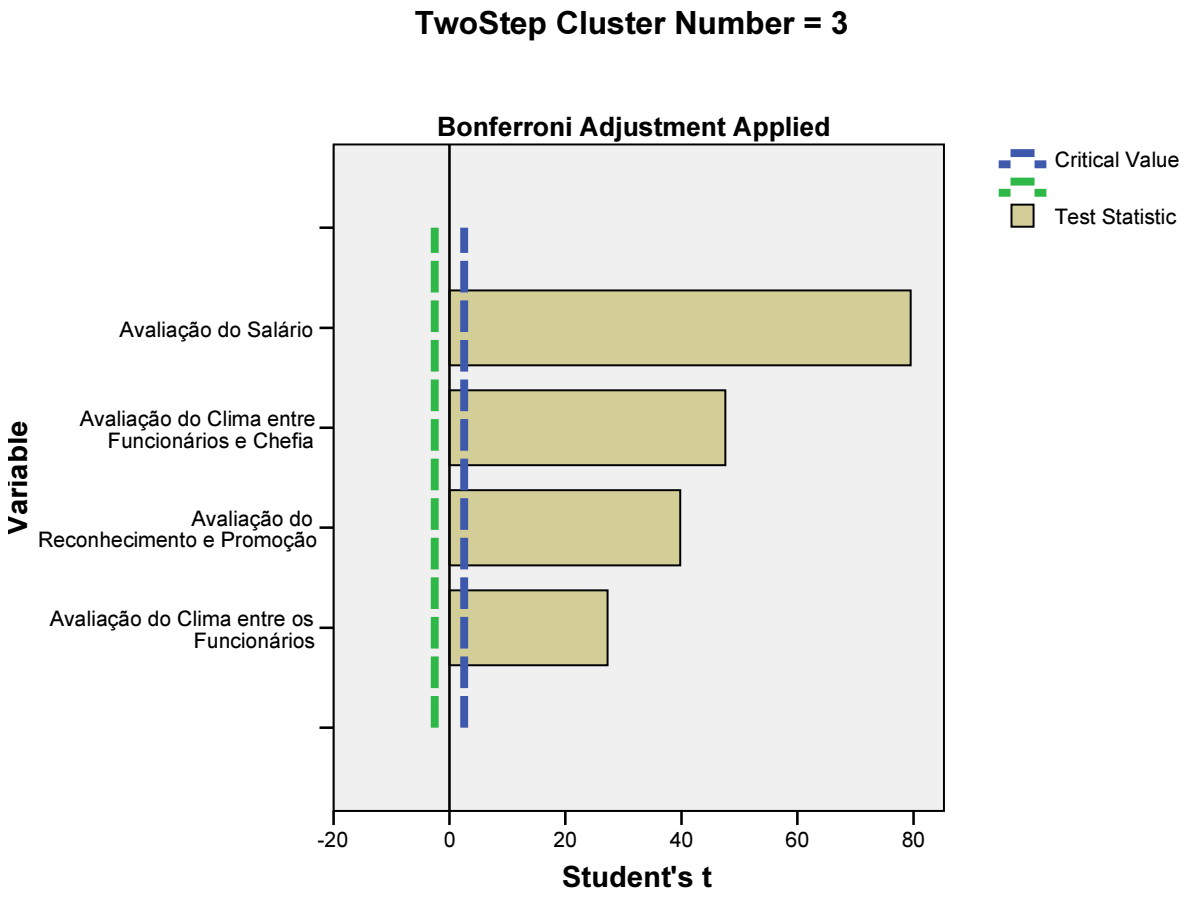


Decisão:

Todas as variáveis da análise são significantes estatisticamente para a formação do cluster 2, isto é, discrimina-o e são importantes para a sua formação.

No Cluster 3:





Decisão:

Todas as variáveis da análise são significantes estatisticamente para a formação do cluster 3, isto é, discrimina-o e são importantes para a sua formação.

5-Conclusão

Com este trabalho, apresentaram-se as potencialidades da Two Step Cluster e procurou-se sensibilizar o pesquisador para a sua ampla possibilidade de utilização em pesquisas quantitativas, em diversas áreas do conhecimento. Com a Two Step Cluster, pesquisas que envolvam uma multiplicidade de indicadores qualitativos e quantitativos, trabalhados simultaneamente, podem sair do terreno analítico puramente trivial, com uma alternativa avançada e útil para análise de cluster, com uma mineração de dados refinada, que permite o reconhecimento de padrões, antes latentes num espaço de partida.

A consecução da Two Step Cluster é operacionalmente possível graças a sua implementação amigável em pacote estatístico, o SPSS. Este software gera clusters com medidas aritméticas e visuais, que permitem a segmentação da base de dados e, uma análise topológica associada a tipologia de classificação.

Na Two Step Cluster, as estatísticas são geradas por médias e seus valores finais vêm de um processo de escalonamento ótimo. As localizações dos objetos nos clusters são

estabelecidas através de medidas de distância, que neste caso são métricas de “probabilidade log” e “euclidiano”.

Efetuada a rodada da Two Step Cluster, o analista terá como output uma tabela de distribuição de frequência de cada clusters, uma tabela com as médias assumidas pelas variáveis quantitativas do modelo em cada clusters(quando o estudo tiver variáveis quantitativas), distribuição de frequência das categorias de cada variável categórica do estudo em cada cluster(quando o estudo tiver variável qualitativa), gráfico de “pizza” das freqüências dos clusters, gráfico de barras para percentagens dentro dos clusters para as variáveis qualitativas, gráfico dos intervalos de confiança para as variáveis quantitativas, gráfico de significância do qui-quadrado para as variáveis qualitativas do modelo e finalmente gráfico de significância t-Student para as variáveis quantitativas.

Os recursos gerados fornecem informações do número de objetos que foram combinados em clusters, o tamanho relativo de cada clusters, o perfil de cada clusters e a significância da diferença dos clusters.

Com a constatação dos resultados pode ser realizada uma análise de reconhecimento de padrões, de mineração de dados, de segmentação da base de dados, de estruturação do espaço de partida, de investigação de subperfis que se distanciem do perfil geral médio.

A opção privilegiada pela técnica da Two Step Cluster ao estudo de caso se justificou principalmente pela constatação da existência de indicadores de configuração complexa, com a existência de variáveis com métricas distintas, umas de natureza categorial e outras de natureza quantitativa, coexistentes num mesmo universo e que seria justificável a análise de clusters Two Step, para que se realizasse uma análise estrutural pertinente, com mais nitidez, simplicidade e racionalidade. O esforço analítico se justificou pela importância estratégica que a informação provinda da análise se apresenta do estudo de caso em foco.

Concluindo, está proposto uma alternativa consistente para pesquisas em negócios, que se precisem realizar análises exploratórias e estruturais de registros com variáveis classificadas em diferentes tipos de medição, com vistas a tomadas de decisão.

A nossa expectativa é que esta poderosa técnica original de análise multivariada se torne cada vez mais habitual no campo da investigação científica, em várias áreas do saber, com dados estatísticos, em contextos em que é ainda inóspita.

Referências

[1]BANFIELD J. D. and A. E. RAFTERY. (1993). Model-based Gaussian and non-Gaussian clustering.Biometrics, 49. p. 803–821.

[2]FRALEY C. and A.E. RAFTERY. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal, 4. p. 578–588.

[3]FRALEY, C. (1998). Algorithms for model-based Gaussian hierarchical clustering.SIAM Journal on Scientific Computing, 20. p. 270–281.

[4]HUANG, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2. p. 283–304.

[5]KAUFMAN, L. and P.J. ROUSSEEUW. (1990). Finding groups in data: An introduction to cluster analysis. Wiley, New York.

[6]MELIA, M. and D. HECKERMAN. (1998). An experimental comparison of several clustering and initialization methods. Microsoft Research Technical Report MSR-TR-98-06.

[7]THEODORIDIS, S. and K. KOUTROUMBAS. (1999). Pattern recognition. Academic Press, New York. Zhang,

[8]T., R. RAMAKRISHNON and M. LIVNY. (1996). BIRCH: An efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD Conference on Management of Data.p. 103–114, Montreal, Canada..