



APLICAÇÃO DE MINERAÇÃO DE DADOS NAS TRANSAÇÕES DE COMPRAS  
EM EMPRESA DO SEGMENTO DE PETRÓLEO E GÁS.

Carlos Vinícius dos Santos Ninho

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken.

Rio de Janeiro

Março de 2011

APLICAÇÃO DE MINERAÇÃO DE DADOS NAS TRANSAÇÕES DE  
COMPRAS EM EMPRESA DO SEGMENTO DE PETRÓLEO E GÁS

Carlos Vinícius dos Santos Ninho

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO  
LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA  
(COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE  
DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE  
EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

---

Prof. Nelson Francisco Favilla Ebecken, DSc.

---

Prof. Beatriz de Souza Leite Pires de Lima, DSc.

---

Prof. Elton Fernandes, Ph.D.

RIO DE JANEIRO, RJ – BRASIL.

MARÇO DE 2011

Ninho, Carlos Vinícius dos Santos

Aplicação de mineração de dados nas transações de compras em empresa do segmento de petróleo e gás / Carlos Vinícius dos Santos Ninho – Rio de Janeiro: UFRJ/COPPE, 2011.

X, 114 p.: il.; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2011.

Referências Bibliográficas: p. 100-103.

1. Mineração de dados. 2. Análise de agrupamentos. 3. Estratégia aprendizacional. 4. Sistemas integrados de gestão empresarial. 5. Tomada de decisão. I. Ebecken, Nelson Francisco Favilla. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

## **Agradecimentos**

Gostaria de agradecer a todos que de certa forma contribuíram para a realização deste trabalho.

Meu agradecimento especial vai para meu orientador Nelson F. F. Ebecken por sempre ter mantido minha linha de raciocínio no rumo certo.

Em especial também agradeço ao Maurício Onoda, Roberto Harkovsky e Paulo Abreu. Sem o apoio de vocês certamente não teria concluído este estudo. Muito obrigado por tudo.

À minha noiva, companheira e amiga Priscila, por ter compreendido minha ausência durante o período de estudo.

À minha mãe Odete pelo apoio.

Aos amigos de turma, pelas infindáveis discussões durante o curso e troca de informações pela madrugada a fora.

A todos os funcionários do Programa de Engenharia Civil, principalmente para a Egna Castro, que sempre me ajudou bastante nos mais variados assuntos.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

APLICAÇÃO DE MINERAÇÃO DE DADOS NAS TRANSAÇÕES DE COMPRAS  
EM EMPRESA DO SEGMENTO DE PETRÓLEO E GÁS

Carlos Vinícius dos Santos Ninho

Março/2011

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Este trabalho descreve a metodologia utilizada na aplicação de métodos de mineração de dados no histórico das transações de compras de uma empresa do segmento de petróleo e gás. Os experimentos foram realizados a partir das transações gravadas no sistema ERP corporativo e copiados para um ambiente exclusivo de desenvolvimento. Foi desenvolvido uma ferramenta OLAP baseada no componente PivotCube, que facilitou o processo de seleção de atributos e pré-processamento junto aos especialistas do negócio. Ao final, utilizando o software Clementine, foram aplicadas diversas técnicas de mineração de dados. A análise dos resultados obtidos ofereceu importante suporte na tomada de decisão sobre a reestruturação da metodologia de trabalho do setor.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

EVALUATION OF DATA MINING TASKS APPLIED TO HISTORICAL  
PURCHASE ORDERS OF AN OIL & GAS COMPANY

Carlos Vinícius dos Santos Ninho

March/2011

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This paper presents the methodology used to evaluate data mining tasks applied to historical purchase orders from an oil company. The experiments were carried out from the transactions recorded on the corporate ERP system and copied to a unique environment for development. We developed an OLAP tool based on the *PivotCube* component, which facilitated the process of attribute selection and pre-processing together with the business experts. Finally, using the Clementine software, several techniques have been applied to data mining. The results analysis offered strong support in decision making to the restructuring of the work methodology at Purchase department.

# Índice

Agradecimentos .....	iv
Resumo .....	v
Abstract.....	vi
Índice .....	vii
Índice de figura.....	viii
Índice de tabela.....	x
1) Introdução.....	1
1.1) Objetivos .....	4
1.2) Metodologia e recursos tecnológicos .....	4
1.3) Organização do trabalho.....	5
2) Revisão de literatura.....	6
2.1) Evolução do setor de compras.....	6
2.2) Análise de processos e otimização de resultados .....	9
2.2.1) Método PDCA.....	15
2.2.2) Estratégia Aprendizacional .....	19
2.3) Sistemas integrados de gestão empresarial .....	22
2.4) Tarefas de mineração de dados .....	26
2.4.1) Análise de agrupamentos .....	27
Algoritmo: K-Means .....	37
Algoritmo: Dois Estágios – Two Step Cluster .....	42
2.4.2) Regras de associação.....	46
Algoritmo: Apriori Node .....	48
Algoritmo: GRI Node.....	51
2.4.3) Classificação por Indução de Árvores de Decisão .....	53
Algoritmo: C5.0 Node .....	55
3) Desenvolvimento da metodologia.....	57
3.1) Coleta e consolidação dos dados.....	57
Tratamento da base de dados.....	62
3.2) Seleção e pré-processamento .....	64
Data Mart Pedido de Compras.....	67
Data Mart Materiais.....	71
3.3) Processamento – Mineração de dados.....	74
Algoritmo: Two-step cluster.....	75
Algoritmo: K-means .....	78
Algoritmo: GRI Node.....	84
Algoritmo: C5.0 Node .....	86
4) Análise dos resultados.....	93
5) Conclusão .....	98
6) Referências Bibliográficas .....	100
Anexo I – Comandos SQL .....	104
Anexo II – Two-Step Cluster results .....	109
Anexo III – K-means: Representação gráfica dos clusters (k=2 e k=4).....	112
Anexo IV – K-means: Convergência da clusterização .....	113
Anexo V – GRI: Data Mart Materiais .....	114

## Índice de figura

Figura 1 - Faturamento x Valor Comprado .....	2
Figura 2 - Quantidade de Pedidos de Compras x Faturamento .....	3
Figura 3 - Modelo de compras de Kraljic, adaptado de CANIELS et al. (2005). .....	3
Figura 4 - Fluxograma do processo de compras. ....	9
Figura 5 - Divisão de trabalho no setor de compras – elaborada pelo autor. ....	10
Figura 6 - Gráfico de PARETO, elaborado pelo autor. ....	13
Figura 7 - Etapas do ciclo PDCA, adaptado de SILVA (2006). ....	15
Figura 8 - Rampa de melhoria, adaptado de ANDRADE (2003). ....	18
Figura 9 - Estratégia Aprendizizacional – adaptado de BARCELLOS (2004). ....	20
Figura 10 – Evolução dos sistemas ERP – adaptada de COLANGELO (2001). ....	22
Figura 11 - Módulos de um sistema ERP .....	24
Figura 12 - Knowledge Discovery in Database (KDD) adaptada de BISHOP (2006)...	26
Figura 13 - Exemplo de clusterização hierárquica representada por Dendrograma .....	29
Figura 14 - Distância Euclidiana .....	31
Figura 15 - Comparação de medidas de Similaridade .....	32
Figura 16 - Quantificação da similaridade/dissimilaridade entre os objetos “A” e “B”	33
Figura 17 - Definição do número de centróides. ....	37
Figura 18 - Agrupamento com os centróides iniciais. ....	38
Figura 19 - Cálculo de novos centróides. ....	38
Figura 20 - Agrupamento com os novos centróides. ....	39
Figura 21 - Convergência do algoritmo de clusterização. ....	39
Figura 22 - Parâmetros básicos de configuração do k-means no software SPSS. ....	40
Figura 23 - Parâmetros avançados de configuração do k-means no software SPSS. ....	41
Figura 24 - Parâmetros do TwoStep Cluster - SPSS .....	43
Figura 25 - Apriori node - SPSS. ....	48
Figura 26 - Parâmetros de ajuste - Apriori node – SPSS. ....	49
Figura 27 - GRI Node - SPSS. ....	51
Figura 28 - Exemplo de árvore de decisão. ....	53
Figura 29 - C5.0 Clementine .....	55
Figura 30 - Estrutura de replicação de banco de dados. ....	57
Figura 31 – PivotCube .....	58
Figura 32 - Ambiente de desenvolvimento. ....	59
Figura 33 - Modelo Entidade Relacionamento da tabela QUERY_CUBE .....	60
Figura 34 - Variáveis “Fornecedor Pc”, “Cfop”, “UF” e “Comprador” .....	66
Figura 35 - Estrutura de dados. ....	67
Figura 36 - volume de pedidos em função do tempo .....	68
Figura 37 - Two-step cluster – DM MATERIAIS .....	75
Figura 38 - Two-step cluster – DM MATERIAIS – real x normalizado .....	76
Figura 39 - Two-step cluster – DM MATERIAIS – clusters .....	77
Figura 40 - Two-step cluster – DM MATERIAIS – clusters (normalizado). ....	77
Figura 41 - k-means – DM MATERIAIS – k=2 e k=3 .....	78
Figura 42 - k-means – DM MATERIAIS – k=4 e k=5 .....	79
Figura 43 - DM MATERIAIS – comparação entre k-means e two-step para k=2. ....	81
Figura 44 - k-means – DM MATERIAIS: k=3 .....	82
Figura 45 - k-means – DM MATERIAIS: k=3 .....	83
Figura 46- DM Materiais – parâmetros de ajuste do algoritmo GRI Node. ....	85
Figura 47 - Parâmetros de ajuste do algoritmo C5.0 Node. ....	87

Figura 48 - DM REQ COMPRAS – C5.0 Node - conjunto de regras geradas para as variáveis <i>UnidNegocio</i> e <i>Comprador</i> .	87
Figura 49 - DM REQ COMPRAS – C5.0 Node - árvore de decisão gerada para as variáveis <i>UnidNegocios</i> e <i>Comprador</i> .	88
Figura 50 - DM REQ COMPRAS – C5.0 Node - conjunto de regras geradas para as variáveis <i>Comprador</i> e <i>prazoemissãopedido</i> .	89
Figura 51 - DM REQ COMPRAS – C5.0 Node - árvore de decisão gerada para as variáveis <i>Comprador</i> e <i>prazoemissãopedido</i> .	90
Figura 52 - DM REQ COMPRAS – C5.0 Node - conjunto de regras geradas para as variáveis <i>Comprador</i> e <i>SomadeVL_TOTAL_ITEM</i> .	91
Figura 53 - DM REQ COMPRAS – C5.0 Node - árvore de decisão gerada para as variáveis <i>Comprador</i> e <i>SomadeVL_TOTAL_ITEM</i> .	92
Figura 54 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=2)	94
Figura 55 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=3)	94
Figura 56 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=4)	94
Figura 57 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=5)	94

## Índice de tabela

Tabela 1 – Itens de controle, adaptado de SILVA (2006).....	13
Tabela 2 – O método 5W2H, extraído de NEVES (2007). .....	16
Tabela 3 – O método FCA, extraído de NEVES (2007). .....	16
Tabela 4- Coeficientes de similaridade que consideram a ausência conjunta.....	34
Tabela 5 - Coeficientes de similaridade que desconsideram a ausência conjunta.....	35
Tabela 6 - Coeficientes de associação contidos no intervalo [-1,+1] .....	35
Tabela 7 - Exemplo de Regras de Associação – GRI Node .....	52
Tabela 8 - PivotCube: passagem de parâmetros para as stored procedures .....	60
Tabela 9 - Variáveis envolvidas nas transações de compras .....	61
Tabela 10 - Atributos do Data Mart Pedido de Compras .....	67
Tabela 11 - Variáveis e sua dimensionalidade .....	68
Tabela 12 - Tipos de frete.....	69
Tabela 13 - Atributos do Data Mart Materiais .....	71
Tabela 14 - Data Marts: REQ COMPRAS e Materiais .....	73
Tabela 15 - DM Materiais – variáveis utilizadas na clusterização .....	74
Tabela 16 - Two-step cluster – DM MATERIAIS - Mean.....	76
Tabela 17 - Two-step cluster – DM MATERIAIS - Standard Deviation.....	76
Tabela 18- k-means – DM MATERIAIS: k=3 e k=4.....	80
Tabela 19 - k-means – DM MATERIAIS: k=5 .....	80
Tabela 20 - k-means – DM MATERIAIS: quantidade de registros por agrupamento...	81
Tabela 21 - k-means – DM MATERIAIS: percentual de registros por agrupamento....	83
Tabela 22- DM Materiais – variáveis utilizadas pelo algoritmo GRI Node.....	84
Tabela 23 - DM Materiais – parâmetros de ajuste do algoritmo GRI Node. ....	84
Tabela 24 - DM Materiais: Regras extraídas pelo algoritmo GRI Node.....	85
Tabela 25- Grupos de valores negociados por requisição .....	97

# 1) Introdução

O gerenciamento dos negócios no mercado corporativo atual é uma tarefa extremamente complicada. A diminuição das fronteiras, neste caso, causada principalmente pelo processo de globalização e expansão das redes de computadores, aumenta a competitividade<sup>1</sup>, pois a interatividade entre clientes e fornecedores é extremamente estimulada devido à facilidade de comunicação entre as partes, independentemente de sua localização física.

Segundo FERRAZ et al. (1997), o aumento da competitividade faz com que as empresas busquem vantagens frente ao mercado. Estas vantagens podem ser obtidas de várias maneiras, como por exemplo, (i) através da qualidade dos produtos em relação aos concorrentes, (ii) da eficiência no processo de produção, (iii) do relacionamento com o cliente, (iv) da disponibilidade de produtos ou serviços, (v) da diminuição dos prazos de entrega, e, principalmente, (vi) do preço de venda (que muitas das vezes torna-se fator crítico para o sucesso de um negócio).

Em todos estes fatores existe a influência direta ou indireta dos fornecedores. Eles são agentes importantes para o desenvolvimento de vantagem competitiva, conforme pode ser visto em BRAGA (2006). Por isso, desenvolver o relacionamento e a visão de futuro juntamente com os parceiros de negócio da cadeia de suprimentos de sua organização se torna fundamental para implementar estratégias concorrenciais de sucesso.

PORTER (1990) ressalta que o relacionamento entre clientes e fornecedores é influenciado pela forma como é feita a organização interna da empresa cliente, pelos seus canais de distribuição e, principalmente, pelos agentes responsáveis pelo processo de compra. Estes elementos compõem o chamado sistema de valores onde cada um influencia no desempenho geral da empresa, e, conseqüentemente, no desenvolvimento de sua estratégia de vantagem competitiva.

A formação de alianças estratégicas entre empresas permite o compartilhamento de conhecimentos específicos e o aumento de competitividade global, além de reduzir consideravelmente seus custos operacionais, o que reflete diretamente em melhores preços de venda. Fica subentendido então que a atuação dos agentes responsáveis pelos processos de compra influencia diretamente os resultados da corporação.

GOLDRAT et al. (1997) coloca que todo esforço estratégico deve se concentrar na necessidade de maximizar os resultados da empresa, pois a lucratividade é uma condição necessária à sobrevivência de qualquer negócio capitalizado. O resultado, ou lucro, pode ser otimizado a partir do desenvolvimento de estratégias operacionais que foquem na melhoria dos seguintes índices: Faturamento, Inventário e Despesa operacional.

O faturamento (ou ganho), ainda segundo GOLDRAT et al. (1997), indica quanto dinheiro se gera através das vendas realizadas. Por sua vez, o inventário é o

---

<sup>1</sup> Ser competitivo é ter capacidade de desenvolver e implementar estratégias concorrenciais, que permitam ampliar ou conservar uma posição sustentável no mundo dos negócios FERRAZ et al. (1997).

investimento feito na compra de produtos que serão utilizados nas vendas. A despesa operacional é todo o dinheiro consumido com a finalidade de transformar o inventário em ganho, como por exemplo, a compra de consumíveis.

Para trabalhar as questões fundamentadas em PORTER (1990) e GOLDRAT et al. (1997), em relação a atuação dos compradores da corporação e a maximização dos resultados da empresa, utilizaremos como cenário os dados de uma empresa multinacional atuante no segmento petrolífero que está presente em mais de 100 países, dentre eles o Brasil, e possui aproximadamente 800 bases e 43 mil funcionários.

Segundo as informações fornecidas pela empresa durante o período de mobilização de um grande contrato na cidade de Macaé, estado do Rio de Janeiro, é possível quantificar a importância do departamento de compras ao comparar o volume financeiro das transações feitas pela unidade Macaé com seu faturamento. Na Figura 1 e Figura 2 podemos visualizar a expressividade dos valores comprados, porém não temos como relacionar com o bem adquirido (se material de consumo, ativo fixo ou contratação de serviço).

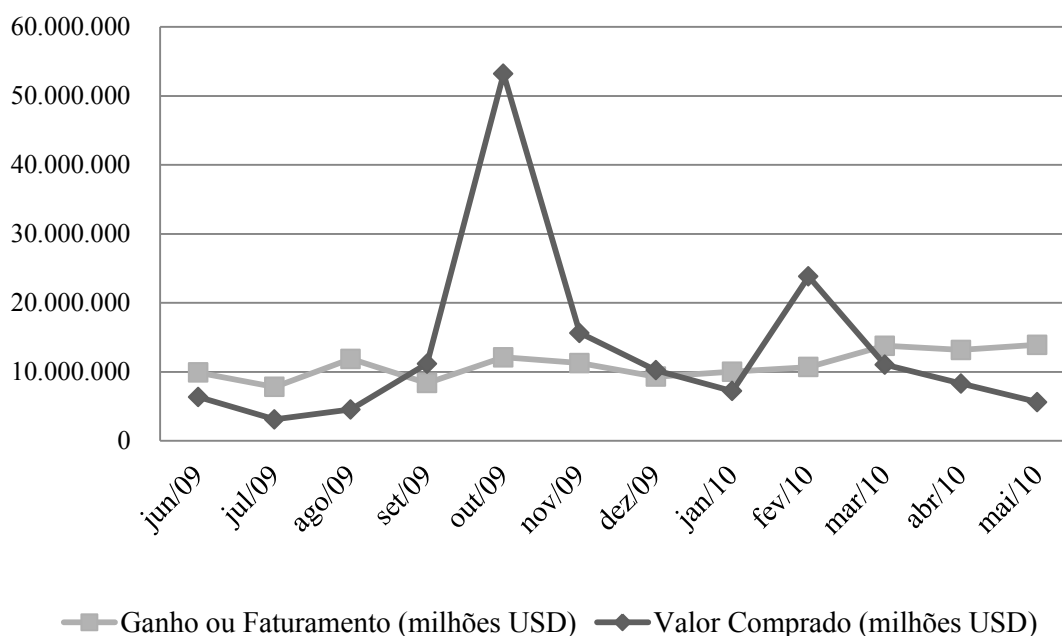


Figura 1 - Faturamento x Valor Comprado

Em BRAGA (2010) vemos que a compra de bens e serviços que são utilizados na produção e na revenda de produtos é o fator de maior contribuição nos custos de produção e das mercadorias vendidas, ou seja, conforme dito em seu artigo, aproximadamente de 50% a 60% do valor final de um produto é repassado para os fornecedores externos. O que evidencia ainda mais a atividade do setor de compras nas organizações.

Vale ressaltar que as figuras demonstram apenas as quantidades e os valores gastos mensalmente. Não temos como informar, neste momento, o que seria despesa operacional ou inventário. De qualquer forma, fica a observação de CHRISTOPHER (1997) e MOORI et al. (2002) em relação à importância da diminuição do estoque de

produtos em manutenção ou acabados, pois evita que a empresa tenha capital parado em inventário, mas fica evidente a expressividade dos valores.

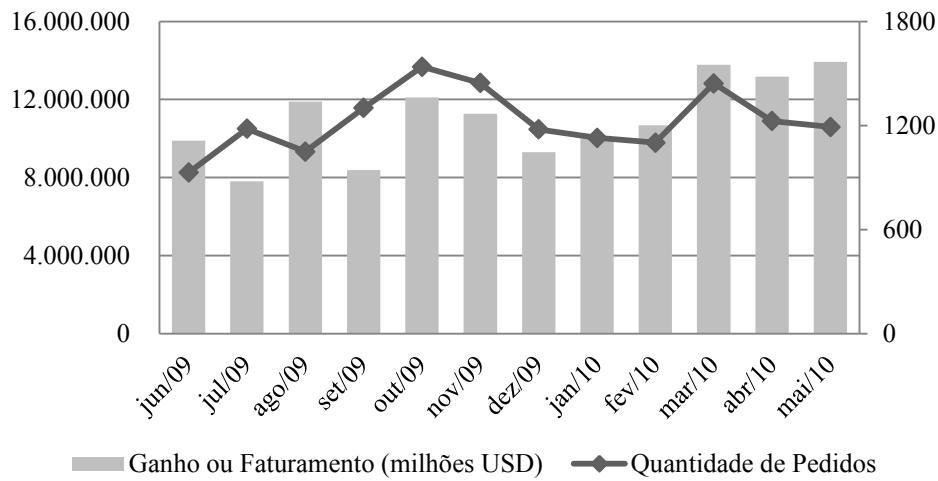


Figura 2 - Quantidade de Pedidos de Compras x Faturamento

Os números apresentados demonstram que o departamento de compras, se bem otimizado, pode colaborar estrategicamente na organização. Para isso, é necessário haver procedimentos, indicadores, rotinas e ferramentas que possibilitem a gestão e a tomada de decisão.

Conforme visto em CANIELS et al. (2005), o modelo de compras de Kraljic (1983) também pode auxiliar no suporte à seleção estratégica de compras, diferenciando produtos por tipos distintos na organização. Este modelo consiste em otimizar a relação entre custos (diretos e indiretos) e risco. A matriz cruza duas dimensões: impacto sobre o resultado financeiro e incerteza de oferta, gerando quatro quadrantes para a categorização de produtos, conforme apresentado na Figura 3.

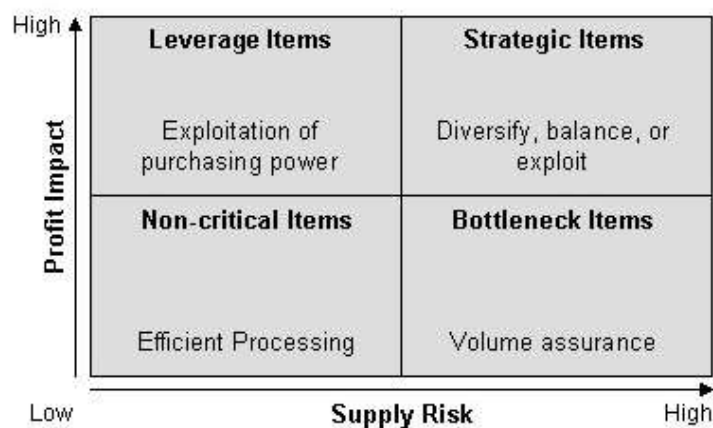


Figura 3 - Modelo de compras de Kraljic, adaptado de CANIELS et al. (2005).

Este modelo permite o foco na administração de compras da organização, evitando a “economia de palitos” (HAVE et al. 2003) e dá margem para que se obtenham vantagens estratégicas pela gestão de fornecedores.

Na empresa utilizada neste estudo, por exemplo, a ferramenta de suporte utilizada é o sistema ERP<sup>2</sup> Sispro<sup>3</sup>. Para auxiliar os gestores não existe nenhuma ferramenta de apoio a decisão disponível, somente os relatórios limitados do sistema ERP.

Baseando-se neste contexto, apresentaremos a seguir os objetivos detalhados desta pesquisa, na qual seus resultados orientarão análises, conclusões e auxiliarão possíveis tomadas de decisão, além da possibilidade de trazer retorno financeiro para a empresa.

## **1.1) Objetivos**

O objetivo deste estudo é pesquisar sobre a importância do departamento de compras na organização e investigar juntamente com os especialistas do negócio por métodos de melhoria de processos. Além disso, espera-se aplicar e avaliar técnicas de mineração no banco de dados no módulo de compras do ERP corporativo na busca por padrões de similaridades entre transações de aquisição de bens de consumo.

Os resultados obtidos servirão de embasamento para uma proposta de readequação na metodologia de trabalho e estratégia de compras da equipe, contribuindo na otimização do setor e no apoio a redução de custos.

Ao final deste estudo, teremos como propor o uso sistemático de uma ferramenta de manipulação direta dos dados corporativos no contexto da Inteligência nos Negócios (*Business Intelligence* - BI), para uso diário dos funcionários do setor de Compras.

## **1.2) Metodologia e recursos tecnológicos**

A estratégia de pesquisa utilizada neste trabalho será baseada no método descritivo<sup>4</sup> exploratório<sup>5</sup> com estudo de caso, buscando nas referências bibliográficas e observações em campo caracterizar o processo existente e propor melhorias nos fluxos pertinentes às atividades de compras.

Na parte experimental, será criado um ambiente de desenvolvimento específico que receberá uma cópia do banco de dados do Módulo de Compras do Sistema ERP Sispro. Aplicaremos ferramentas de mineração de dados que agruparão os registros de acordo com a similaridade entre eles, levando em consideração apenas as variáveis mais

---

<sup>2</sup> ERP (Enterprise Resource Planning) ou SIGE (Sistemas Integrados de Gestão Empresarial) são ferramentas computacionais que integram todos os dados e processos de uma organização em um único sistema, LAUDON et al. (2004).

<sup>3</sup> Maiores informações em <http://www.sispro.com.br>.

<sup>4</sup> Segundo GIL (1991), a pesquisa descritiva visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis. Envolve o uso de técnicas padronizadas de coleta de dados: questionário e observação sistemática.

<sup>5</sup> GIL (1991) também descreve que a pesquisa exploratória proporciona maior familiaridade com o problema com vistas a torná-lo explícito ou a construir hipóteses. Envolve levantamento bibliográfico, entrevistas com pessoas que tiveram experiências práticas com o problema pesquisado e análise de exemplos que estimulem a compreensão.

pertinentes. Utilizaremos também ferramentas de extração de regras de associação e árvores de decisão. Após o descobrimento dos padrões faremos a análise em conjunto com o especialista visando o desenvolvimento de uma ferramenta que permita a manipulação direta dos dados da produção de forma multidimensional.

Todos os experimentos serão feitos utilizando os recursos disponíveis nos softwares: Microsoft SQL Server 2005, Visual Basic, Delphi, Clementine, Matlab, Excel e o ERP Sispro.

### **1.3) Organização do trabalho**

Este trabalho está dividido em seis capítulos.

No primeiro capítulo foi apresentado o contexto em que este trabalho está inserido e sua metodologia de pesquisa.

No segundo capítulo é feita a revisão bibliográfica, que embasa os resultados obtidos no planejamento estratégico do setor de Compras. Neste momento, é discutido as diversas fases de amadurecimento do departamento em questão bem como algumas ferramentas utilizadas no gerenciamento e otimização de processos. Além de fazermos a análise exploratória das atividades de compras, também é feita a revisão bibliográfica das técnicas de mineração de dados e os algoritmos que serão utilizados na fase de desenvolvimento.

O terceiro capítulo descreve toda a atividade prática desenvolvida no processo de busca de conhecimento em banco de dados. Neste capítulo temos o detalhamento de toda a etapa de coleta e consolidação dos dados, seleção e pré-processamento e a aplicação de algoritmos de mineração de dados.

Todos os resultados obtidos são demonstrados no quarto capítulo. Logo em seguida, no quinto capítulo temos as conclusões finais.

As referências bibliográficas podem ser vistas no sexto, e último, capítulo.

## 2) Revisão de literatura

Neste capítulo discutiremos sobre a importância estratégica do setor de compras e veremos algumas metodologias de análise e melhoria de processos. Faremos também a revisão sobre as técnicas de *data mining* utilizadas na parte experimental do trabalho.

### 2.1) Evolução do setor de compras

Conforme citado anteriormente no capítulo um, e visto em BRAGA (2006) e MONCZKA et al. (2003), um dos maiores componentes do custo de produção de mercadorias vendidas está associado à aquisição de bens e serviços, ultrapassando, em muitos casos, 50% do valor do produto. Estes custos estão associados à compra de equipamentos, ferramentas, matéria-prima e contratação de serviços terceirizados, sendo esta a responsabilidade do setor de compras.

O setor de compras passa a se tornar estratégico ao negócio a partir do momento que a alta direção compreende sua evolução no decorrer do tempo. BRAGA (2006) coloca que o desenvolvimento de Compras nos últimos dez anos pode ser compilado em quatro fases distintas bastante comuns em muitas organizações.

Na primeira fase os clientes internos, ou requisitantes, fazem todo o processo de cotação, negociação de preços, condições de pagamento e datas de entrega, restando ao setor realizar apenas a parte burocrática da compra, que consiste no cadastro do fornecedor e emissão dos pedidos. Ou seja, apenas executam os acordos firmados por terceiros.

Neste momento percebe-se a falta de visão compartilhada entre o setor e toda a corporação, detalhado por SENGE (1998), que acaba desestimulando o comprometimento e deixando de lado a oportunidade de criar uma forte sinergia entre os setores da empresa na busca da otimização dos resultados; pois o trabalho dos compradores é reativo. BRAGA (2006) coloca que seu desempenho está relacionado simplesmente à capacidade de emissão de documentos e controle de pedidos. Atuam como soldados obedientes realizando tarefas metódicas, combatendo uma grande pilha de transações que aguardam serem despachadas.

Identificamos, juntamente com o usuário especialista, que a empresa utilizada nesta pesquisa se encontra na segunda fase, onde há preocupação com as despesas. Todas as negociações são feitas através do setor de Compras por pessoas especializadas em determinadas aquisições. Neste ponto o grupo de compradores possui um objetivo comum: atingir a meta - reduzir custos, que permite fomentar o sentimento de coletividade no setor. A visão compartilhada muda a relação das pessoas com a companhia, elas se sentem parte de um todo.

A comunicação no departamento evolui, pois buscam o melhor entendimento possível sobre as necessidades dos clientes internos. Uma importante disciplina é desenvolvida: aprendizagem em grupo, exigindo a prática do diálogo e da discussão, conforme colocado por SENGE (1998). Os compradores analisam questões complexas sob

diferentes pontos de vista, comunicando suas idéias e discutindo livremente com os requisitantes todos os pontos pertinentes ao negócio. A cada nova transação o conhecimento coletivo cresce e os problemas relacionados à compra de suprimentos diminuem, mas o trabalho ainda é feito de forma mecânica.

É fácil encontrar nesta fase os limites do crescimento do departamento. O grupo de trabalho se aperfeiçoa, devido à grande interação com os demais setores e gera bons frutos (capital salvo para a organização) durante algum tempo e depois para de melhorar. Os resultados do setor se estabilizam. Não há alinhamento do setor com as estratégias competitivas da empresa.

Ainda em BRAGA (2006) vemos algumas características importantes desta fase, como a forma na qual é medido o desempenho do setor, que é baseada estritamente na redução de custos e a percepção da alta direção, que reconhece o valor de seus compradores, mas ainda não enxerga claramente como eles podem agregar valor estratégico a empresa. Nota-se a influência do setor nos resultados financeiros, mas falta proatividade em relação a problemas operacionais.

A terceira fase compreende uma pequena variação da anterior. Já existe um grande envolvimento do setor com assuntos operacionais de médio e longo prazo e outros departamentos são envolvidos para garantir o sucesso das aquisições.

O trabalho mecânico da fase anterior é identificado e desenvolve-se a disciplina do pensamento sistêmico na equipe de Compras, que passa a identificar padrões de comportamento entre fornecedores. Esta disciplina possui enfoque especial em SENGE (1998).

Nos casos onde os sistemas de estoque *just-in-time*<sup>6</sup> são empregados, a relação de confiança entre clientes e fornecedores não se sustenta, visto que os fornecedores acabam exigindo exclusividade para compensar o risco de atender o fabricante de um dia para o outro, pois precisam manter altos níveis de inventário. Isso ameaça a equipe de compras do fabricante, que está acostumada a negociar menores preços com diversos outros fornecedores.

Temos o surgimento de equipes multifuncionais, que são times formados por compradores e especialistas de diversas áreas da companhia que trabalham em conjunto na avaliação de fornecedores para o fechamento de contratos de venda de suprimentos e contratação de serviços.

O comprador é valorizado por sua experiência em negociação de assuntos estratégicos da empresa.

Na última fase, conforme visto em BRAGA (2006), temos o setor de Compras completamente integrado com a empresa. Atua diretamente na construção do

---

<sup>6</sup> O sistema *Just-in-Time* (JIT), também conhecido como o “Sistema Toyota de Produção”, é uma filosofia de administração da manufatura idealizada pela Toyota no Japão, nos meados da década de 60, que busca vantagem competitiva através da otimização do processo produtivo sem necessidade de qualquer estoque adicional, seja na expectativa de demanda futura, seja como resultado de ineficiência no processo de fabricação, MOURA et al. (1994).

planejamento estratégico e relaciona-se diretamente com a alta gerência, pois faz parte dela. Segundo OLIVEIRA (2001), os três tipos de planejamento utilizados como ferramenta de definição de metas são:

- (i) Estratégico – desenvolvido pela alta gerência;
- (ii) Tático – desenvolvido pela média gerência e
- (iii) Operacional – desempenhado pela equipe de chão de fábrica.

O elemento-chave para obter sucesso em seu planejamento estratégico, independente do nível hierárquico, é desenvolver um sistema de alinhamento de todos os funcionários em torno desse mesmo objetivo.

*“Quando o funcionário sabe a razão de estar fazendo o que lhe foi incumbido, o realiza de forma mais consciente e satisfeita, pois se sente parte do processo”.* SENGE (1998).

A estratégia de uma organização deve refletir seus planos de desenvolvimento de valor para seus acionistas, clientes e cidadãos. Seja através de seus planos financeiros para aumento do faturamento e redução de despesas, ou baseados na conquista de novos clientes de produtos e serviços. KAPLAN et al. (2004) diz que a estratégia também pode contemplar atividades como melhoria de processos, controle ambiental e recursos humanos.

MONCZKA (1998) afirma que a concentração de esforços para atender as necessidades operacionais, apoiar a área de engenharia, contribuir na redução de custos e no aumento da qualidade de produtos e serviços, deve focar na otimização do gerenciamento de fornecedores e efetivação de contratos de fornecimento que assegurem o fornecimento de suprimentos em tempo hábil para a produção e operação.

As vendas, assim como os pagamentos, devem ser programadas e alinhadas com as devidas compras. O departamento de Compras deve conhecer o Lead-Time (tempo de reposição) do seu fornecedor e emitir os pedidos de acordo com esse tempo. O dinheiro da empresa não pode ficar parado em inventários sem fim. Estoque Zero não significa deixar faltar, quer dizer zero de perda com produtos armazenados. Evitando desperdiçar o capital de giro com manutenção e limpeza do local de armazenamento de produtos (que poderia ser menor), IPTU relativo a este espaço, controle, contagem e segurança da mercadoria, possibilidade de obsolescência, sem levar em consideração que este dinheiro poderia ser aplicado e render juros.

Crescer o estoque nem sempre é crescimento da empresa. O setor de Compras deve estar alinhado a este fundamento e agir de forma proativa no desenvolvimento de menores tempos de reposição.

Os papéis estratégicos que o setor de Compras representa, de acordo com PEARSON (1999), incluem o fato de colher informações do mercado buscando novas soluções, produtos e serviços, previsão de preços, planejamento de longo prazo e determinação de políticas de Compras e realizar a análise constante de valor dos itens adquiridos.

Como não é o escopo deste trabalho tratar das responsabilidades do departamento de Compras, não detalharemos o mapa de processos e matriz de responsabilidade e, sim, focaremos na análise de processos e busca de otimização de tarefas.

## 2.2) Análise de processos e otimização de resultados

Depois de tomado conhecimento do cenário que envolve esta pesquisa, devemos nos ambientar aos principais métodos e ferramentas utilizadas no mercado que buscam a otimização de resultados.

OLIVEIRA (2001) diz que para praticar melhorias em procedimentos operacionais, é fundamental, em primeiro lugar, realizar a análise dos processos, que garantirá ao gestor ou responsável pela melhoria, que todo o fluxo de informações e interações de processos é conhecido. Além disso, a análise aprofundada permitirá a adequação do fluxo das operações às pessoas que as executam, identificando a necessidade de treinamentos específicos e, até mesmo, verificar as vantagens em alterar a seqüência das atividades.

Segundo OLIVEIRA (2002), a documentação dos processos na maioria dos casos é reproduzida através de fluxograma, uma representação gráfica que apresenta a seqüência de um trabalho de forma analítica, caracterizando as operações, os responsáveis e/ou unidades organizacionais envolvidos no processo.

A elaboração de fluxogramas de análise de processos com a representação gráfica de cada passo de uma atividade facilita a criação de indicadores de desempenho e ajuda na identificação de pontos de interação de atividades de setores distintos.

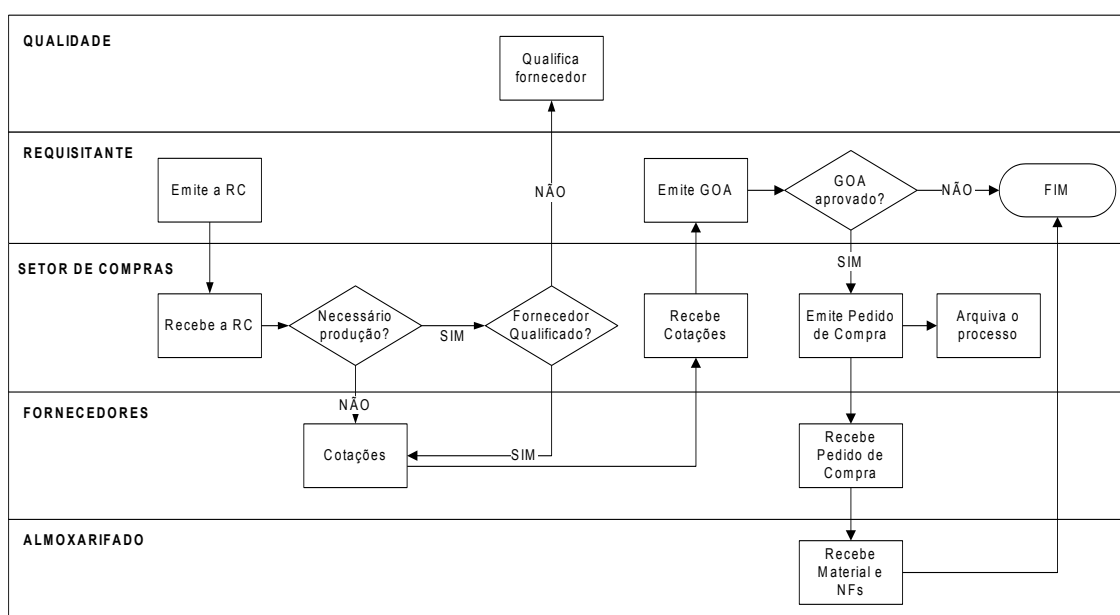


Figura 4 - Fluxograma do processo de compras.

Obtêm-se aumento da eficiência quando se resolve questões de tarefas que não estão claramente definidas, principalmente quando tratamos de definição de responsabilidades.

A Figura 4, extraída do repositório de procedimentos internos da empresa em questão, apresenta o fluxograma da operação diária do setor de Compras da empresa utilizada nesta pesquisa.

Numa leitura simplificada, quando o setor de Compras recebe uma requisição de compra (RC), é verificado junto ao requisitante da necessidade de elaboração fabril do material. Se for preciso e não tivermos um fornecedor qualificado para a fabricação do produto, o departamento de Controle de Qualidade é acionado para homologar novos fornecedores juntamente com o comprador e um representante técnico indicado pelo requisitante.

A partir deste momento a equipe está apta para dar início ao processo de cotação junto aos fornecedores homologados.

Como a empresa em questão lida com vários segmentos (chamado também de unidades de negócio ou linhas de produto) na área de Petróleo e Gás, cada comprador do departamento é responsável pelo atendimento de determinadas linhas de produto, Figura 5. A divisão de trabalho foi estipulada baseada na quantidade de RC's emitidas, ou seja, as linhas consideradas "carro chefe da empresa" possuem compradores exclusivos, enquanto que unidades de negócio menores compartilham um único comprador.

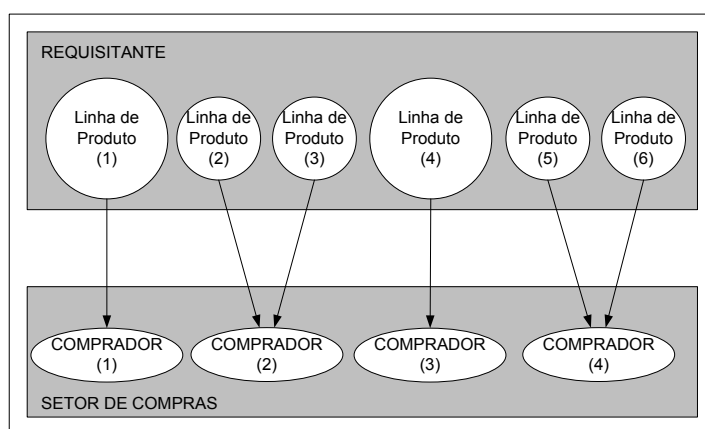


Figura 5 - Divisão de trabalho no setor de compras – elaborada pelo autor.

Esta metodologia de trabalho foi implementada para evitar atrasos nas transações de compra e funciona de forma satisfatória em relação a atendimento de prazos de entrega de produtos, mas não leva em consideração o tipo de material ou serviço contratado. É comum termos o mesmo material/equipamento sendo comprado em diferentes fornecedores porque o processo foi feito por compradores distintos.

Em cada transação é necessário pelo menos três cotações, dependendo do valor a ser negociado, que são encaminhadas para o solicitante que fica responsável pela obtenção da aprovação financeira via sistema GOA (*Grant Of Authority*) para efetivação da compra.

O processo de aprovação financeira, identificado como gargalo no processo, é um sistema localizado na intranet usado para todas as aprovações de compras de materiais ou serviços. Ao submeter uma ordem de pagamento, o pedido é enviado para uma cadeia de aprovação que varia dependendo do valor e natureza do processo. O tempo médio de aprovação varia de quatro dias a duas semanas. Nenhum pagamento é efetuado sem a documentação aprovada neste sistema.

Com o valor aprovado o pedido é emitido ao fornecedor. Todas as cotações e documentos pertinentes a transação são arquivados.

O recebimento é feito pelo Almoxarifado, que após ter o material inspecionado finaliza o processo informando ao requisitante que o produto está disponível para retirada.

Embora implícito, o fluxo dos dados atribuídos às atividades do setor não é representado na figura acima. Estas informações são documentadas em datagramas, que, segundo OLIVEIRA (2001), por mais parecidos com os fluxogramas que sejam, possuem enfoque diferente. Priorizam o fluxo dos dados e não do processo. São muito utilizados por programadores de sistemas na especificação de novos produtos.

A análise aprofundada de processos feita em Compras foi registrada num documento titulado Procedimento Operacional de Compras (POC), e, disponibilizado para consulta somente na intranet da companhia.

O objetivo do documento é estabelecer padrões, registros e responsabilidades no procedimento de compras, assegurando a aquisição de produtos no prazo requerido, melhores condições de custo, qualidade em atendimento às especificações técnicas corretas e homologação dos fornecedores. As ressalvas sobre os pontos mais relevantes do Procedimento Operacional de Compras estão descritos a seguir.

## PONTOS IMPORTANTES

Durante os encontros de preparação do POC contamos com a participação de diversos compradores e clientes internos. Na última fase de redação, tivemos a revisão do procedimento feita pelo "*Supply Chain Coordinator*" - coordenador da cadeia de suprimentos e validação do "*Internal Audit Manager*" - gerente de auditoria interna.

Depois de concluído, o documento foi disponibilizado para consulta mediante aprovação da alta gerência do Brasil, composta pelo "*Country Controller*" - gerente financeiro, "*Country Supply Chain Manager*" - gerente da cadeia de suprimentos e o "*Country Manager*" - gerente geral, níveis hierárquico mais alto no país.

As normas e responsabilidades que foram estabelecidas aplicam-se a todos os compradores e funcionários da empresa, principalmente no caso do requisitante, que é a pessoa responsável pela (i) identificação da necessidade da aquisição do bem ou serviço, (ii) emissão da Requisição de Compra, (iii) obtenção da aprovação através da assinatura manual ou eletrônica do seu supervisor e/ou gerente e (iv) submeter o processo para aprovação via sistema GOA.

Os funcionários da empresa que emitem requisições de compras não estão autorizados a emitir os Pedidos para os fornecedores ou nem efetuar compras pessoais em nome da empresa.

Cabe ao setor de Compras realizar todo e qualquer contato com fornecedores, solicitando apoio e participação, quando necessário, das áreas técnica e financeira no processo de negociação. Além de realizar toda a negociação e acompanhamento dos pedidos, o departamento mantém atualizada a base de dados cadastrais de identificação dos fornecedores. Nenhum fornecedor pode ser cadastrado ou homologado sem passar pelo setor.

Ao receber uma cópia da RC gerada pelo requisitante, o Setor de Compras verificará o correto preenchimento e assinatura da supervisão e/ou gerência. Caso a documentação (Requisição de Compra assinada, GOA aprovado e cotações) esteja em conforme, dará andamento a emissão do Pedido de Compra. Se a documentação não estiver em conformidades, será devolvida ao requisitante para as correções que se fizerem necessária.

Quanto à quantidade de cotações, deve-se obedecer aos níveis de valores estabelecidos abaixo:

- ✓ 1 (uma) Cotação: Compras até US\$ 100.00 (cem dólares).
- ✓ 2 (duas) Cotações: Compras maiores que US\$ 100.00 (cem dólares) e menores que US\$ 400.00 (quatrocentos dólares).
- ✓ 3 (três) Cotações: Compras superiores a US\$ 400.00 (quatrocentos dólares).

A negociação deverá ser realizada sempre com o fornecedor que atender os seguintes critérios de avaliação:

- 1º) A melhor Qualidade
- 2º) O melhor preço
- 3º) O melhor prazo de entrega
- 4º) Outros (condições de pagamento, transporte, atendimento, etc.)

A Qualificação e Cadastro de Fornecedores têm como objetivo compor, mediante critérios específicos pré-definidos, uma base de fornecedores homologados para fornecimento de materiais e serviços.

O Almoxarifado deve fazer o recebimento de materiais e equipamentos, que devem ser encaminhados para a inspeção no setor de Controle de Qualidade sempre que aplicável.

Por fim, o Controle de Qualidade deve conferir e liberar os produtos em conformidade com as especificações técnicas inseridas no escopo da compra. Caso seja verificada alguma não conformidade no produto/serviço, o inspetor designado deverá registrar o ocorrido no Relatório de Não Conformidade via sistema encaminhando-o ao setor de Compras para que o fornecedor seja notificado e tome as devidas providências.

Segundo NEVES (2007), com o detalhamento das atividades a alta gerência pode definir itens de controle e estipular metas para as atividades mais relevantes. Estes controles estabelecem grandezas mensuráveis, tais como tamanho e quantidade de atrasos.

Na Tabela 1 temos um exemplo de como é possível traduzir algumas características do trabalho em indicadores de desempenho ou de qualidade.

Característica	Indicador	Fórmula
Negociações com fornecedores	Índice de Preços	$\frac{\text{Soma mensal dos preços reais pagos}}{\text{Soma mensal dos preços médios de mercado}}$
Satisfação com o trabalho	Índice de Faltas	$\frac{\text{Faltas não justificadas no mês}}{\text{Total de faltas no setor}}$
Custo Operacional	Custo Operacional por requisição de compras	$\frac{\text{Custo operacional mensal total do setor de Compras}}{\text{Quantidade de RC's atendidas}}$

Tabela 1 – Itens de controle, adaptado de SILVA (2006).

Segundo KUME (1993) e SANTOS et al. (2005), as análises dos indicadores podem ser feitas de forma ágil e eficaz quando se utiliza a representação gráfica, principalmente o gráfico de PARETO, que é uma forma especial do gráfico de barras verticais, que dispõe os itens analisados desde o mais freqüente até o menos freqüente.

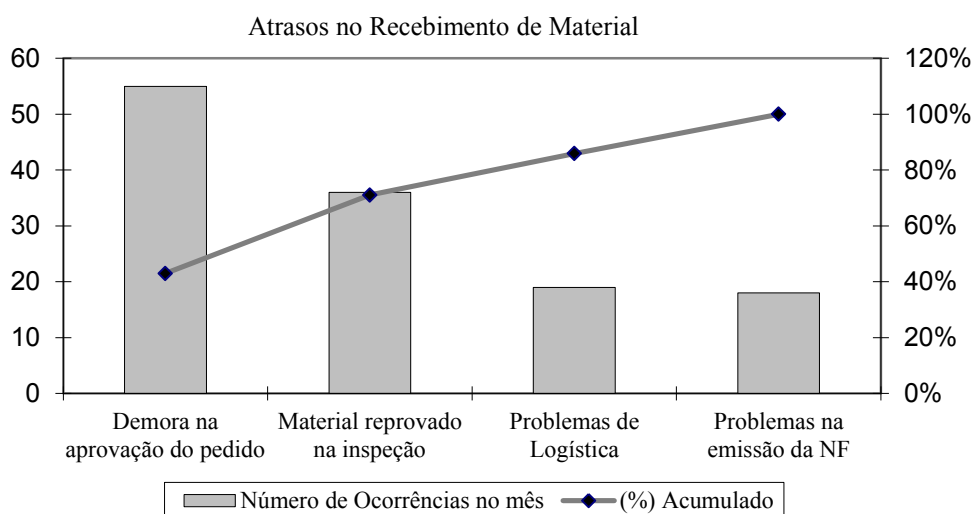


Figura 6 - Gráfico de PARETO, elaborado pelo autor.

O gráfico de Pareto foi desenvolvido pelo economista e sociólogo italiano Wilfredo Frederico Samaso, que viveu entre os anos 1848 e 1923, e, tem como objetivo estabelecer prioridades na tomada de decisão. A partir de uma abordagem estatística é possível avaliar quais são os problemas mais graves e que precisam ser tratados com mais urgência, VIEIRA (1999).

Neste estudo de caso, o Pareto exemplificado na Figura 6 realça de forma bem próxima a realidade os maiores causadores de atrasos no recebimento de materiais: a demora na aprovação financeira para efetivar a compra e reprovação do material na inspeção feita pelo Controle de Qualidade. Representando aproximadamente 80% dos motivos de atraso de liberação de material para o cliente interno.

Segundo VIEIRA (1999), na análise do gráfico de Pareto devem-se considerar os seguintes pontos importantes:

- ✓ Não tomar decisões baseado num único levantamento de dados, ou conclusão de um Pareto, pois a curva de criticidade pode ter sido influenciada por um fenômeno isolado ocorrido durante a coleta de informações. Faça dois ou mais levantamentos e compare as curvas de desempenho.
- ✓ Buscar a alta gerência para definir que problemas deverão ser tratados como prioridade: os mais frequentes ou os mais caros. Pode acontecer de um problema corriqueiro não trazer prejuízo significativo para a companhia.

Identificado a(s) dificuldade(s), é necessário fazer uma análise para apurar as causas e os efeitos.

*“... sempre que um grande número de causas contribui para um determinado efeito, poucas dessas causas são as responsáveis pela maior parte dos efeitos.” SILVA (2006).*

Mostraremos a seguir alguns métodos e ferramentas utilizadas na pesquisa de causas e efeitos de problemas em processos operacionais.

## 2.2.1) Método PDCA

Em SOUZA (1997), vimos que a prática do método, ou ciclo, PDCA (do inglês Plan, Do, Check, Act – Planejar, Fazer, Checar e Agir) como ferramenta de controle de qualidade começou a ser divulgado a partir da década de 50 por William Edwards Deming no Japão. É um método seqüencial de análise e solução de problemas, que formarão um ciclo de melhoria contínua para o alcance de metas.

Cada etapa do ciclo, representado na Figura 7, tem um objetivo distinto e dependente da etapa anterior, exceto quando é a primeira vez que se implementa o método:

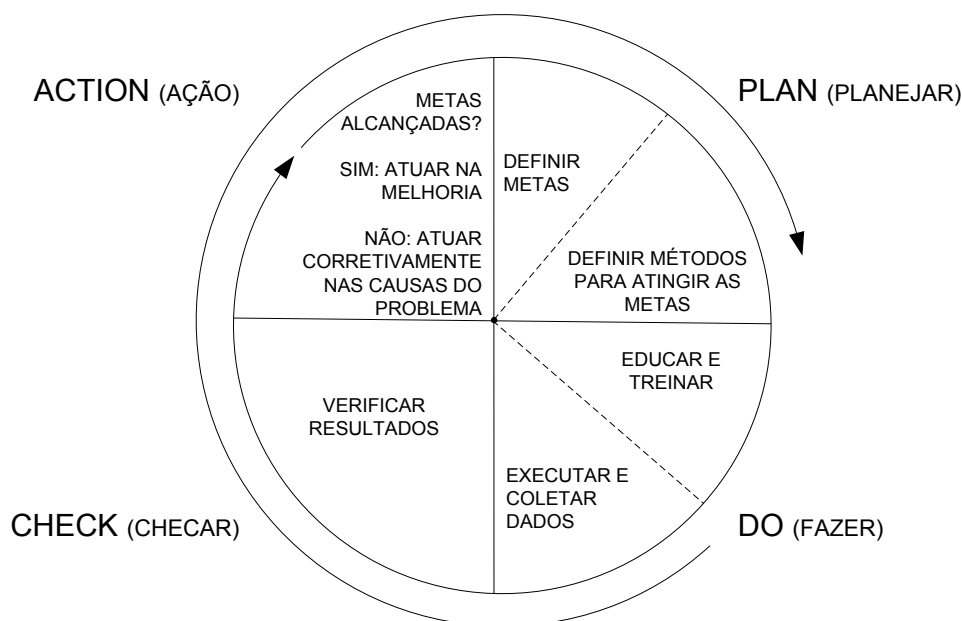


Figura 7 - Etapas do ciclo PDCA, adaptado de SILVA (2006).

O início desta metodologia começa no planejamento (**PLAN**) estratégico que definirá as metas e os recursos que serão utilizados para atingi-las. Esta é a fase mais importante, pois é o início do ciclo que desencadeará todo o processo de melhoria contínua do PDCA.

ANDRADE (2003) ressalta que todos os envolvidos, principalmente a alta administração, devem procurar meios para melhorar seus negócios e atividades diárias, participando os funcionários, alocando os recursos e estimando prazos para a efetivação do plano de ação.

Os dados que serão coletados para a base de indicadores de evolução de processos são definidos nesta etapa. Segundo CAMPOS (1992), para que todas as premissas sejam atendidas, devem-se seguir os seguintes passos:

- (i) Localizar o problema,
- (ii) Instituir metas,
- (iii) Analisar as causas e, por último,
- (iv) Definir plano de ação.

O plano de ação deve conter prazos para todos os resultados aguardados na fase de planejamento. No caso de algum resultado não estar de acordo com o esperado, podemos utilizar as ferramentas 5W2H e FCA para o tratamento dos desvios dos indicadores.

Segundo KAPLAN et al. (2004), tanto o método FCA quanto o 5W2H<sup>7</sup>, surgiram para facilitar a identificação das variáveis de um processo, garantindo que todos os ângulos sejam abordados.

De acordo com OLIVEIRA (2002), citado por SANTOS et al. (2005), estes métodos

*“devem ser estruturados para permitir uma rápida identificação dos elementos necessários à implantação do projeto”.*

Os elementos envolvidos no caso 5W2H são:

<b>5W</b>	What	O que?	O que será executado?
	When	Quando?	Quando a ação será executada?
	Who	Quem?	Quem é o responsável pela ação?
	Where	Onde?	Onde será executada a ação?
	Why	Por Quê?	Por que a ação será executada?
<b>2H</b>	How	Como?	Como será executada a ação?
	How much	Quanto custa?	Quanto custa para executa a ação?

Tabela 2 – O método 5W2H, extraído de NEVES (2007).

FCA significa Fato-Causa-Ação e foi criado para auxiliar no gerenciamento de não conformidades, além de atender as exigências do órgão certificador ISO 9001:2000. Sua estrutura é bastante simples:

Fato	Causa	Causa	Causa	Ação	Responsável
Descrição da não-conformidade	1º Possível Causa	2º Possível Causa	3º Possível Causa	Ações para solucionar o problema	Responsável pelas ações.

Tabela 3 – O método FCA, extraído de NEVES (2007).

Ao abrir um FCA o responsável deverá concluir com as ações dentro do prazo estipulado durante o planejamento.

Com todos os objetivos traçados e devidamente documentados em um plano de ação, chega-se a próxima etapa do ciclo, definida como FAZER (**DO**), ou EXECUTAR.

<sup>7</sup> Segundo DJAIR (2008), a recomendação do uso do 5W2H é bastante antiga, tendo seu registro mais antigo encontrado no "Tratado sobre Oratória", escrito por Marcus Fabius Quintilianus entre os anos 30 e 100 D.C.

Segundo CAMPOS (1992), o sucesso dessa fase depende basicamente de duas atividades: treinamento e ação. No treinamento a empresa deverá divulgar o plano de ação para todos os funcionários, fomentando a disciplina de visão compartilhada, SENGE (1998), e buscando o comprometimento de todos no cumprimento das metas. Os treinamentos deverão apresentar as atividades que deverão ser executadas, as formas como a evolução do trabalho será medida e os seus respectivos prazos.

Devem-se manter constantes as verificações na empresa durante a execução das atividades, a fim de esclarecer possíveis dúvidas durante a execução do serviço. Todos os resultados deverão ser registrados, sejam positivos ou negativos, para que sejam avaliados na próxima do ciclo.

Nas datas estipuladas é feito a verificação dos resultados (**CHECK**) da fase anterior e comparação com a meta planejada. É de suma importância o suporte de uma metodologia estatística para que se minimize a possibilidade de erros e haja economia de tempo e recursos. A análise dos dados desta fase indicará se o processo está de acordo com o planejado.

ANDRADE (2003) coloca que na quarta e última etapa do ciclo (**ACTION**) é preciso agir em relação aos resultados analisados na fase anterior. Para isso, encaramos as seguintes situações:

1) *A meta foi atingida:* neste caso os resultados devem ser anunciados e o plano proposto documentado para que as mudanças feitas sejam padronizadas. Uma data deve ser definida como marco inicial para a nova metodologia operacional, a fim de evitar confusões nos setores envolvidos.

A partir deste momento sugere-se a implantação do gerenciamento da rotina de CAMPOS (2001), que concentra seus esforços na eliminação de não conformidades decorrentes da variação e integração entre processos. Segundo NEVES (2007) alguns autores sugerem a metodologia SDCA, recorrente do PDCA sendo que o 'S' estaria relacionado à padronização (STANDARD), para manter a padronização de processos.

2) *A meta não foi cumprida ou foi parcialmente atingida:* deve-se iniciar outro ciclo baseado nas ações que foram tomadas anteriormente que não surtiram os efeitos esperados.

O processo de melhoria contínua se dá quando terminamos um ciclo PDCA e logo depois iniciamos outro, partindo de um novo patamar de qualidade de processos e buscando resultados ainda melhores.

O aprendizado obtido da conclusão dos ciclos permite, que seja feito avanços na solução de problemas cada vez mais complexos, criando uma rampa de melhoria, conforme podemos observar na Figura 8. Segundo CAMPOS (1992), isso é melhoria contínua no sentido de Qualidade Total.

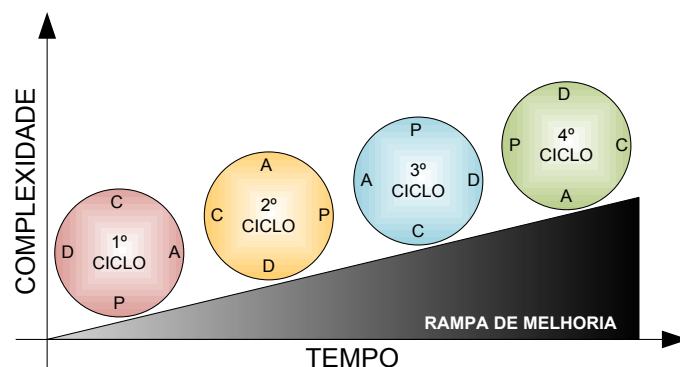


Figura 8 - Rampa de melhoria, adaptado de ANDRADE (2003)

Algumas das dificuldades encontradas em implementar a melhoria contínua dos processos – fazer o ciclo PDCA girar – estão na falta de comprometimento com cada etapa.

Na fase de planejamento (**PLAN**), por exemplo, os esforços somam 120%. Existe um excesso de vontade no planejamento das metas enquanto que os esforços para atingi-las (**DO**) não chegam a 80%, principalmente quando é necessário investimento. Com isso, a fase de verificação dos resultados (**CHECK**) é feita pela metade, em virtude das pendências da etapa anterior. Alguns administradores nem chegam a consolidar as metas (**ACTION**). Dão o plano de trabalho como encerrado assim que notam a melhora no processo.

Mesmo assim, empresas estão obtendo grandes retornos com o uso desta metodologia, aprendendo a analisar e solucionar os problemas, adquirindo maior satisfação dos clientes e evitando os retrabalhos e desperdícios.

Como dito por Sun Tzu em seu famoso livro A arte da Guerra:

*“Se conhecemos o inimigo e a nós mesmos, não precisamos temer uma centena de combates. Se nos conhecemos, mas não ao inimigo, para cada vitória sofreremos uma derrota. Se não nos conhecemos nem ao inimigo, perderemos todas as batalhas.”*  
Sun Tzu - A Arte da Guerra.

O PDCA se mostra muito mais eficiente quando aplicado numa organização que sabe identificar seus pontos fortes e fracos.

## 2.2.2) Estratégia Aprendizacional

O último método de otimização de resultados e cumprimento de metas abordada nesta pesquisa é apresentado sob o conceito de Estratégia Aprendizacional, defendido por Paulo Barcellos em 2004 no Rio de Janeiro. Segundo sua própria definição:

*“A estratégia aprendizacional é um neologismo no qual a aprendizagem é a base e o veículo para o contínuo desenvolvimento da estratégia, que por sua vez estará cada vez mais voltada para a promoção da aprendizagem, formando um círculo virtuoso, composto pelo aprendizado da estratégia e pela estratégia voltada para o aprendizado”.*  
BARCELLOS (2004).

Para o desenvolvimento do aprendizado estratégico é indispensável um ambiente que promova estas práticas, conforme esquema apresentado na Figura 9, que correlacione os **quatro conceitos** chaves desta metodologia:

- (i) Percepção, Criatividade e Participação;
- (ii) Mapas Estratégicos e Base Unificada de Conhecimentos Estratégicos;
- (iii) Planejamento Estratégico Participativo Aprendizacional e
- (iv) Sistemas Inteligentes.

O **primeiro conceito** é composto pela Percepção, Criatividade e Participação, que são os princípios norteadores de sustentação da metodologia.

De acordo com BETHLEM (1996), o comportamento das pessoas é baseado na interpretação que fazem da realidade e não na realidade em si. Por este motivo, a percepção do mundo é diferente para cada um de nós. Cada pessoa percebe um objeto ou uma situação de acordo com os aspectos que têm especial importância para si própria. De certa forma, podemos descrever percepção como a forma contínua que enxergamos o que há ao nosso redor. Uma nova descoberta pode ser encarada como uma reação criativa a um estímulo da percepção (através de testes e observações).

A criatividade é o elemento capaz de transformar uma ameaça em oportunidade. Mas para isso acontecer dentro das empresas, é importante que invistam em ambientes que promovam a criatividade de seus colaboradores, parceiros, clientes e fornecedores. Segundo GODFREY citado por BARCELLOS (2004), estes ambientes devem estar imunes à burocracia, eliminando as barreiras físicas entre departamentos, facilitando a transmissão de novas ideias entre diferentes níveis hierárquicos e, conseqüentemente, aumentando a participação de todos na estratégia da organização. BARCELLOS (2004) propõe ainda a participação democrática de todos os envolvidos no processo estratégico, pois como terão a incumbência de executar a estratégia, é preciso que todos a entendam.

No **segundo conceito** da estratégia aprendizacional encontramos os mapas estratégicos que definem um modelo de correlação dos indicadores.

De acordo com KAPLAN et al. (2004), não se consegue gerenciar o que não se pode medir, e não se consegue medir o que não se pode descrever. O mapa estratégico tenta

resolver este problema, ao fornecer um modelo para uma representação gráfica simples, numa única página, das relações de causa e efeito entre os objetivos tanto das dimensões crescimento e processos internos, quanto das dimensões mercadológica e econômico-financeira da estratégia.

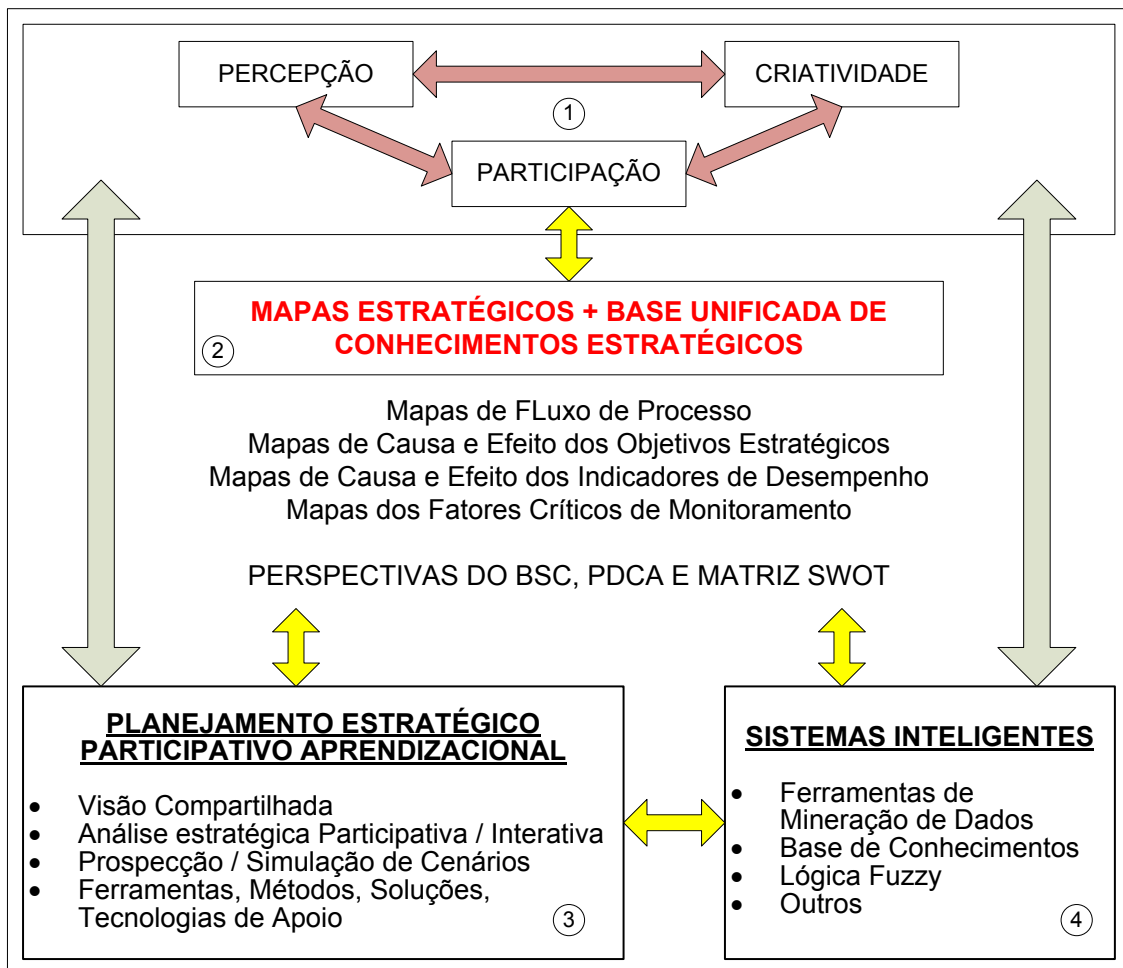


Figura 9 - Estratégia Aprendizacional – adaptado de BARCELLOS (2004).

Os mapas estratégicos são utilizados em conjunto com a Base Unificada de Conhecimentos Estratégicos.

A idéia de utilização de uma base de conhecimentos, na qual os conhecimentos (estratégicos) explícitos da organização poderão estar disponibilizados (de forma unificada), para interação dos diversos colaboradores, que, a partir deste processo de interação (com seus conhecimentos tácitos), poderão criar conhecimento novo e explicitá-los nesta base, e assim sucessivamente, formando uma espécie de “espiral do conhecimento”, BARCELLOS (2004).

O Planejamento Estratégico Participativo Aprendizacional, **terceiro conceito** de nossa estratégia, ressalta a importância da percepção e criatividade de todos os funcionários da empresa no processo de planejamento estratégico da organização, permitindo o desenvolvimento do conceito de visão compartilhada, elucidado por SENGE (1998).

Em sua pesquisa, BARCELLOS (2004), viabiliza a participação na gestão estratégica e a aprendizagem dela decorrente através de um ambiente virtual desenvolvido

especialmente para este propósito, que permite o uso de ferramentas e sistemas dinâmicos e inteligentes nos processos de discussão, análise, teste, simulação e monitoramento, da estratégia e dos elementos a ela associados.

Por fim, no **quarto e último conceito**, descrevemos o uso de Sistemas Inteligentes que podem contribuir na tomada de decisão servindo como instrumento de fundamentação técnica para a alta gerência.

*“Os sistemas inteligentes, podem potencializar nossas propostas, e até mesmo viabilizá-las em alguns casos, motivo pelo qual os apresentamos como sendo um dos elementos que se integram para promover um efetivo processo de criação, desenvolvimento e implementação daquilo que passamos a chamar de estratégia aprendizacional, em todas as suas fases e em diversas perspectivas.”*  
BARCELLOS (2004).

Quando se têm disponível todas as informações necessárias, incluindo os dados históricos, as ferramentas de inteligência artificial podem ajudar na definição de objetivos, indicadores e metas através de técnicas de classificação, hierarquização, seleção e priorização. É possível, por exemplo, medir o grau de pertinência destes indicadores utilizando os conceitos da lógica nebulosa (fuzzy), detalhada mais adiante neste mesmo capítulo, e até mesmo utilizar da sumarização ou agrupamento de textos para extrair “contribuições pertinentes” aos problemas da organização, dentre uma massa de sugestões, críticas, comentários etc.

Os dados que manipularemos neste estudo são provenientes de um sistema corporativo conhecido como ERP (do inglês Enterprise Resource Planning). Mostraremos a seguir algumas das características destes sistemas.

### 2.3) Sistemas integrados de gestão empresarial

Segundo PROTIL et al. (2006), as corporações industriais que tiveram maior destaque do mundo deram início, a partir dos anos 90, à adoção dos sistemas integrados de gestão empresariais conhecidos pela sigla ERP (Enterprise Resources Planning). CORRÊA et al. (2001) coloca que estes sistemas são frutos de um processo evolutivo iniciado na década de 60, representado pela Figura 10, que focava as necessidades das empresas manufactureiras na automatização do tratamento das listas de compra de materiais. Para se ter uma idéia de como era complicado realizar esta tarefa, podemos imaginar uma grande fábrica gerenciando listas de materiais em torno de 5.000 a 10.000 itens de estoque por produto final produzido. Se este controle fosse feito hoje, os recursos computacionais disponíveis atenderiam perfeitamente as nossas necessidades, mas não em meados dos anos 50.

A infraestrutura de TI dos anos 60 viabilizou esta automatização e permitiu que coordenasse melhor a compra de suprimentos em função da capacidade de produção das fábricas, o que diminuiu consideravelmente o tamanho dos estoques e, conseqüentemente aumentou o capital de giro.

Em COLANGELO (2001) vimos que com os aprimoramentos das ferramentas computacionais, na década de 70, já era possível prever a capacidade de produção e correlacionar esta capacidade com o consumo de itens de estoque. Surge então uma poderosa ferramenta de gestão chamada de **MRP** (Material Requirements Planning - Planejamento da Necessidade de Materiais).

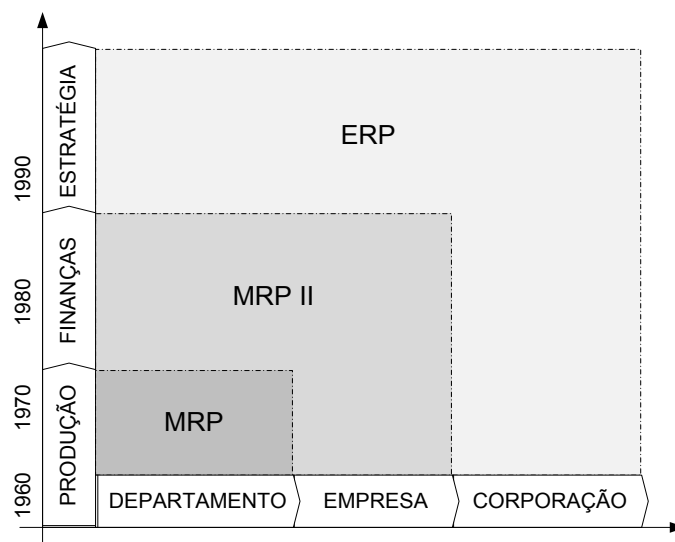


Figura 10 – Evolução dos sistemas ERP – adaptada de COLANGELO (2001).

O MRP, ou planejamento de necessidades de materiais, converte a previsão de demanda em programação da necessidade de utilização de seus suprimentos. De acordo com PROTIL et al. (2006), a partir do conhecimento de todos os componentes de um determinado produto e os tempos de obtenção de cada um deles, podemos calcular no MRP o quanto e quando se deve obter de cada item, de forma que não haja falta e nem sobra do suprimento nas necessidades da produção.

Mas para que todos os dados acima sejam calculados corretamente, alguns dos parâmetros básicos para o funcionamento do MRP precisam ser configurados, conforme sugerido por COLANGELO (2001), como por exemplo:

- ✓ Estrutura do Produto (quantidade de cada item que compõem um produto),
- ✓ Tempo de Reposição (tempo gasto entre a colocação do pedido de compra até o recebimento do material),
- ✓ Tempo de Fabricação e
- ✓ Estoque mínimo e máximo (a quantidade que deve ser mantida em estoque, seja de matéria-prima ou produto acabado).

Importante ressaltar em CARNEIRO (2005) que os sistemas MRP dos anos 70 não suportavam ao planejamento de capacidade produtiva e custos, e não se integravam com outras aplicações usadas pela organização. Limitavam-se ao planejamento de produção, compras e controle de estoque. Estes recursos seriam adicionados posteriormente, de acordo com as prioridades (demandas) estabelecidas pelo mercado.

COLANGELO (2001) e CORRÊA et al. (2001) acrescentam que nos anos 80 foram adicionados ao MRP os módulos *SFC* (Shop Floor Control – controle de fabricação), *CRP* (Capacity Requirements Planning - planejamento de necessidades de capacidade produtiva) e *Purchasing* (controle de compras), mudando sua nomenclatura para o que conhecemos hoje como **MRPII** (Manufacturing Resource Planning – ou planejamento de recursos de manufatura).

Os sistemas MRPII além de executar funções de planejamento de produção e estoque, tratam também do planejamento de capacidades produtivas (como a capacidade de máquina e os recursos humanos necessários) e dos recursos financeiros (como orçamento e custeio da produção), CORRÊA et al. (2001). Durante todos os anos 80, os esforços dos mantenedores de sistemas MRPII estavam focados no aperfeiçoamento da ferramenta. Muitos avanços foram feitos, mas as empresas queixavam-se que o MRPII não estava trazendo os benefícios esperados, pois não estava integrado com os demais softwares da organização (como por exemplo, Vendas, Contábil, Financeiro e Recursos Humanos).

Foi a partir dos anos 90 que tivemos os maiores avanços no MRPII. A tecnologia computacional da época, incluindo as tecnologias de rede e melhorias nas telecomunicações, segundo COLANGELO (2001), permitiu a integração com outros sistemas corporativos, como o administrativo, fiscal, recursos humanos, entre outros. Os desenvolvedores desta solução tiveram que mudar o foco de seu desenvolvimento, que antes se baseava no Planejamento de atividades, para Integração dos dados de diversos setores da organização numa base de dados única. Essa evolução levou ao surgimento dos sistemas hoje conhecidos como **ERP** – Enterprise Resource Planning (ou SIGE - Sistemas Integrados de Gestão Empresarial).

SOUZA (2000) define os sistemas ERP como um conjunto de softwares comerciais que atuam de forma integrada com a finalidade de suportar a maioria das operações de uma empresa. Normalmente são divididos em módulos, conforme visto na Figura 11, que se

comunicam e atualizam a mesma base de dados, de modo que os dados alimentados em um determinado módulo são instantaneamente disponibilizados para os demais.



Figura 11 - Módulos de um sistema ERP

Alguns dos benefícios da utilização de sistemas ERP, citados por SOUZA (2000), estão a forte integração do sistema, que permite o controle da empresa como um todo; a padronização de relatórios e indicadores gerenciais, que apresenta a melhoria da qualidade na informação fornecida pelo sistema e a disponibilização de melhores práticas para o redesenho dos processos internos da empresa.

Os pontos negativos deste tipo de sistema estão ligados, além do alto custo, a grande dificuldade para sua implementação. Em alguns casos conhecidos no mercado, a implementação do ERP chegou a levar 5 anos para ser completado devido a complexidade na readequação dos processos da organização, que perde a visão hierárquica e departamental e passa a ser orientada a processos que cruzam e integram departamentos.

O ERP que será utilizado no experimento desta pesquisa foi desenvolvido pela empresa Sispro, que atua no mercado desde 1972 com bases operacionais no Rio Grande do Sul, São Paulo e Rio de Janeiro. Foi implementado na empresa em questão em janeiro de 2008 e possui em funcionamento os seguintes módulos (maiores informações em [www.sispro.com.br](http://www.sispro.com.br)):

- ✓ Contábil: disponibiliza um completo conjunto de ferramentas (Contabilidade Fiscal, Contabilidade Gerencial, Contabilidade Orçamentária e Livros Fiscais) que, aliadas ao seu elevado grau de parametrização, permite ajustar-se a diferentes tipos de organizações, facilitando também sua integração com outros sistemas de gestão.
- ✓ Finanças: propicia maior eficiência na administração e gerenciamento financeiro. Os processos integrados de Contas a Pagar, Contas a Receber, Tesouraria, Faturamento e Contratos agilizam e apóiam os gestores na tomada de decisão.
- ✓ Patrimônio: proporciona mais eficácia na apropriação dos custos de depreciação, planejamento dos investimentos e controle físico, contábil e fiscal do imobilizado, garantindo total controle sobre seu ativo fixo.

- ✓ Planejamento e Controle de Produção: é uma ferramenta que automatiza e gerencia as rotinas da área industrial nos moldes dos conceitos de MRPII.
- ✓ Recursos Humanos: otimizam os processos de gestão de RH e folha de pagamento com a automatização dos processos de Folha de Pagamento, Cargos e Salários, Treinamento e Desenvolvimento, Ponto Eletrônico, Controle de Benefícios, Saúde Ocupacional e Segurança do Trabalho.
- ✓ Suprimentos: oferece um importante conjunto de soluções que agilizam os processos de Compras, Recebimento e Movimentação de Estoques, possibilitando as melhores práticas de gestão.

Para manter o escopo da pesquisa, focaremos nossos esforços do módulo de Suprimentos, que se aplicados as técnicas de extração de conhecimento disponíveis, pode se transformar em um importante diferencial estratégico para a empresa.

No próximo item abordaremos os conceitos de *data mining* (mineração de dados) que serão aplicados no módulo de Suprimentos do ERP Sispro.

## 2.4) Tarefas de mineração de dados

De acordo com MARKOV et al. (2007), mineração de dados, ou data mining, pode ser explicada como sendo um conjunto de técnicas baseadas em formulações matemáticas que buscam extrair conhecimento de uma grande massa de dados. Faz parte de um processo conhecido como KDD (Knowledge Discovery in Database) ou Busca de Conhecimento em Banco de Dados, exemplificado na Figura 12, que possui basicamente quatro fases:

1. **Consolidação dos dados:** onde diferentes fontes de dados são combinadas produzindo um único repositório de informações. Nesta etapa fazemos uma limpeza nos registros, pois são eliminados ruídos e dados inconsistentes.
2. **Seleção e Pré-processamento:** aqui são selecionados os atributos que realmente interessam ao usuário. Os dados ficam preparados num formato apropriado para aplicação de algoritmos de mineração.
3. **Mineração:** fase principal do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões, tendências e/ou regras de associação inerentes a base de dados.
4. **Interpretação e avaliação (ou Pós-processamento):** aqui são identificados os padrões realmente interessantes entre os diversos obtidos na fase anterior, de acordo com algum critério pré-definido. São utilizadas técnicas de visualização e representação do conhecimento para introduzir os usuários aos resultados alcançados.

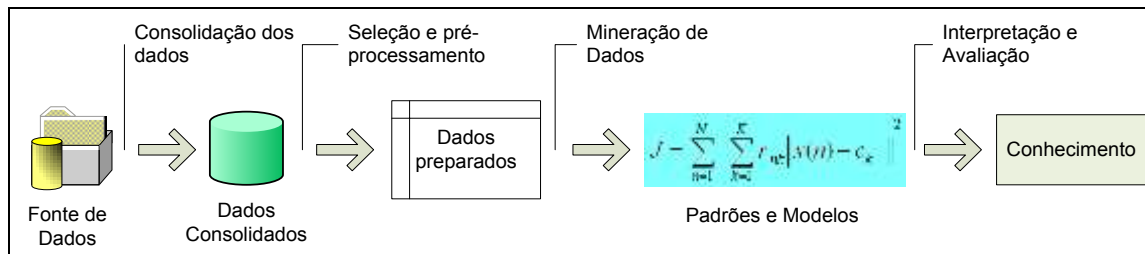


Figura 12 - Knowledge Discovery in Database (KDD) adaptada de BISHOP (2006).

Os padrões interessantes são armazenados como uma nova base de Conhecimento, que poderá ser utilizada como ferramenta de apoio a tomada de decisão.

Segundo HAN et al. (2001), a diferença entre o que vem a ser *técnica* e o que é *tarefa* de mineração de dados é observada pelos seguintes conceitos:

A *tarefa* baseia-se na especificação do que estamos buscando nos dados ou que tipo de padrões seria mais importante. Pode ser subdivididos em tarefas preditivas, que fazem uma inferência a partir dos dados presentes, de maneira a fazer previsões sobre dados futuros e, tarefas descritivas, que caracterizam as propriedades gerais dos dados na base de dados.

As tarefas preditivas mais utilizadas são chamadas de classificação e regressão, enquanto que as descritivas mais comuns são as regras de associação, agrupamento (ou clusterização) e sumarização.

Segundo BISHOP (2006) a *técnica* de mineração consiste na particularização de métodos que nos garantam como descobrir os padrões que nos interessam. Dentre as principais técnicas utilizadas temos, por exemplo, árvores de decisão, métodos de particionamento e análises estatísticas.

Em virtude do foco desta pesquisa utilizaremos técnicas específicas que melhor se adaptam às características da base de dados disponibilizada.

### **2.4.1) Análise de agrupamentos**

O termo Análise de Agrupamentos, também conhecido por clusterização, primeiramente usado por TYRON (1939), está relacionado a diferentes algoritmos de classificação, todos com o objetivo de organizar (categorizar) dados em estruturas (grupos) que façam sentido, sendo estas estruturas baseadas nas características comuns de cada indivíduo.

Para a formação dos grupos (clusters) são utilizadas técnicas estatísticas multivariadas, com conotação exploratória, que verificam a similaridade dos objetos, através de coeficientes específicos para cada tipo de variável - os coeficientes de similaridade, bem como os tipos de variáveis serão vistos em seguida neste mesmo capítulo.

Segundo HAN et al. (2001), algoritmos de clusterização têm sido aplicados em diversas áreas. Na área médica, por exemplo, agrupamento de doenças por sintoma ou curas pode levar a taxonomias muito úteis. Em áreas da psiquiatria, por exemplo, considera-se que o agrupamento de sintomas como paranoia, esquizofrenia e outros é essencial para a terapia adequada. Na arqueologia, por outro lado, também se tem tentado agrupar civilizações ou épocas de civilizações com base em ferramentas de pedra, objetos funerários, etc. De forma geral, toda vez que se faz necessário categorizar (agrupar) dados desconhecidos em classes gerenciáveis, se utiliza métodos de agrupamento.

Repare que no exemplo anterior, o clusterizador não possui informações a priori dos grupos em que os dados devem ser categorizados. Esta é a diferença entre clusterização e classificação: na classificação as regras que definem os grupos são conhecidas a priori. Cabe ao classificador verificar o indivíduo que mais combina com cada classe, tendo como métrica de avaliação a estrutura de cada grupo, HAN et al. (2001). Quando as estruturas de cada classe (ou grupo) não são conhecidas, o agrupamento é feito a partir de técnicas de clusterização, que considera a similaridade (ou dissimilaridade) de cada atributo de um indivíduo.

Segundo PAKHIRA et al. (2004), o resultado obtido a partir da clusterização é um conjunto de grupos com coesão interna e isolamento externo, ou seja, elementos dentro de um mesmo grupo são tão similares quanto possível e são, ao mesmo tempo, tão dissimilares quanto possível dos elementos presentes nos demais grupos. Esta é a problemática da análise de agrupamentos: recebido um conjunto de dados, de objetos, tentar agrupá-los de forma que os elementos que compõem cada grupo sejam mais

parecidos entre si do que parecidos com os elementos dos outros grupos, colocando os iguais (ou quase) juntos num mesmo grupo e os desiguais em grupos distintos.

Conforme visto em HAN et al. (2001) e BISHOP (2006), os métodos de agrupamento existentes são determinados, basicamente, pela medida de proximidade e pelo algoritmo empregado. As medidas de proximidade avaliam a relação entre indivíduos das variáveis observadas e os algoritmos descrevem como o procedimento de agrupamento deve ser realizado. O método ideal deve atender aos seguintes requisitos:

- ✓ Descobrir grupos com forma arbitrária. (Os métodos de Clustering baseados nas medidas de distância Euclidiana ou Manhattan tendem a encontrar clusters esféricos de tamanho e densidade similares).
- ✓ Encontrar o número adequado de clusters. Muitos métodos precisam de um valor de referência.
- ✓ Identificar grupos de tamanhos variados.
- ✓ Aceitar os diversos tipos de variáveis possíveis, por exemplo: escaladas em intervalos, binárias, nominais (categóricas) e ordinais.
- ✓ Ser insensível a ordem de apresentação dos objetos. Neste caso, o algoritmo deverá apresentar os mesmos resultados independentemente da ordem das variáveis.
- ✓ Permitir objetos com qualquer número de atributos (também chamado de variáveis ou dimensões).
- ✓ Ser escalável. Deve permitir lidar com grande quantidade de objetos.
- ✓ Fornecer resultados compreensíveis e utilizáveis.
- ✓ Tolerância a ruídos. A maioria das bases de dados reais contém ruídos ou dados faltantes, o que não deve afetar a qualidade dos clusters obtidos.
- ✓ Aceitar restrições - aplicações reais necessitam agrupar objetos de acordo com vários tipos de restrições. Os métodos devem encontrar grupos de dados com comportamento que satisfaça as restrições especificadas.
- ✓ Manter o mínimo de parâmetros de entrada possíveis.

Conforme visto acima, alguns algoritmos possuem características especiais que precisam ser analisadas com cautela como, por exemplo, a inicialização do algoritmo e a determinação de parâmetros. A medida de proximidade e o algoritmo são objetos que detalharemos ao longo desta pesquisa.

Segundo TYRON (1939), nenhuma técnica de agrupamento atende a todos os pré-requisitos acima simultaneamente, mas um trabalho considerável tem sido feito para atender a cada ponto separadamente.

Segundo HAN et al. (2001), uma classificação mais geral dos algoritmos de clusterização divide os algoritmos em dois métodos distintos:

1. Métodos hierárquicos – envolvem a construção de uma hierarquia em forma de árvore. Há basicamente dois tipos de métodos hierárquicos: aglomerativos e divisivos. Os aglomerativos geralmente começam com um número de clusters iguais ao tamanho do banco de dados (um grupo para cada objeto). A partir da distância entre os objetos, eles são aglutinados a cada iteração até que uma condição de parada seja fornecida ou o número  $k$  (conhecido a priori) de clusters seja alcançado.

O método divisivo inicia um grande cluster agregando todos os objetos que, a cada iteração, vão se dividindo e formando novos clusters. No limite, o número de clusters formado é igual ao tamanho do banco de dados (um grupo para cada objeto).

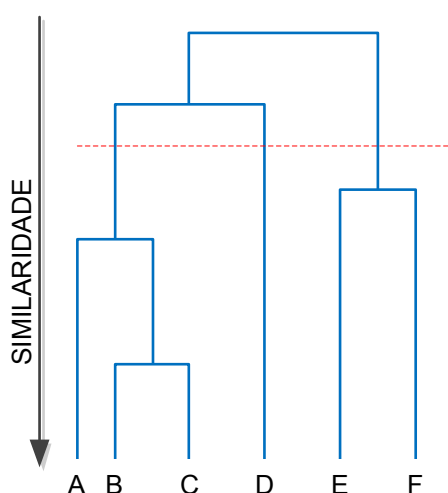


Figura 13 - Exemplo de clusterização hierárquica representada por Dendrograma

Na Figura 13 estes grupos são representados numa estrutura conhecida como dendrograma. Segundo LAPOINTE et al. (1991), um dendrograma é definido como uma árvore com raiz e ponderada, onde todos os nós terminais estão a uma mesma distância (comprimento do caminho) da raiz.

Como podemos ver no exemplo, os elementos B e C pertencem a mesma classe e, juntos, formam um agrupamento em nível superior com o elemento A. Abaixo da linha de corte (parâmetro de avaliação definido pelo usuário especialista) encontramos os elementos agrupados da seguinte forma: um cluster formado pelos elementos {A,B,C}, outro formado por {E,F} e {D} classificado como centro de um único agrupamento.

2. Métodos por particionamento - dividem a base de dados em  $k$  grupos, onde  $k$  é o número de agrupamentos que se deseja formar. Neste tipo de método o usuário informa ao algoritmo o valor de  $k$ . Cada grupo ( $k$ ) possui um centróide  $c(k)$  e os objetos são classificados a partir da menor distância entre cada objeto e  $c(k)$ .

O objetivo dos métodos particionais é encontrar a matriz de coordenadas dos centros de agrupamentos, geralmente calculada pela otimização de uma função objetivo ( $J$ ), muito conhecida como critério do erro quadrático, que será detalhada a seguir quando descrevermos o algoritmo k-means.

Em virtude da diversidade de algoritmos disponíveis, existe grande dificuldade na definição da técnica que deve ser empregada, pois cada uma nos leva a resultados diferentes devido ao fato de os agrupamentos finais serem fortemente dependentes da metodologia usada. Por isso, para não comprometer os resultados obtidos, faremos estudos comparativos entre diversos algoritmos de mineração de dados disponíveis no software SPSS.

Veremos a seguir as estruturas de dados que são suportadas por estes algoritmos e a forma como sua similaridade e/ou dissimilaridade são estabelecidos.

### **Tipos de Variáveis**

Conforme visto em HAN et al. (2001), para que seja possível agrupar dados, é importante definir alguns tipos de variáveis que o algoritmo possa lidar. Como por exemplo:

*Variáveis escaladas em intervalos:* são medidas contínuas que utilizam unidades do de medida do tipo quilograma, litro, metro, etc. A unidade de medida, pode influenciar o algoritmo de análise de agrupamentos. Dependendo da unidade adotada, dois objetos podem ter um valor de similaridade maior ou menor. Como no caso de um determinado atributo ter o valor de 9 metros e outro de 7 - calculando a diferença entre eles temos o valor de 2 metros. Se a unidade de medida fosse centímetros, a diferença seria de 200 centímetros. Por isso, todas as medidas são escaladas para a unidade correta antes de ser aplicada a medida de similaridade entre os objetos.

*Variáveis ordinais:* assemelham-se a uma variável nominal, pois possuem um conjunto finito de valores, mas neste caso, existe uma ordem específica que pode assumir valores discretos ou contínuos. A avaliação da dissimilaridade destas variáveis pode ser feito como as variáveis escaladas – por exemplo, os dias da semana (segunda-feira, terça-feira, etc.), HAN et al. (2001).

*Variáveis nominais (ou categóricas):* possuem um conjunto finito de valores e não possuem uma ordem específica. Exemplo: estado civil - solteiro, casado, viúvo ou divorciado. Por não haver ordenamento, não podemos dizer que "solteiro" é maior que "casado", com isso, não podemos associar medidas numéricas para estas variáveis.

*Variáveis livres:* não possuem estrutura. Texto livre.

*Variáveis binárias ou booleanas:* este tipo de variável pode assumir apenas dois tipos de valores (normalmente 0 ou 1). Quando a variável tem o valor igual a zero significa que o objeto não possui determinada característica, e quando é igual a 1, que ele a possui. Tratar valores binários como valores numéricos pode levar a análises de clusters errôneas, por isso, as dissimilaridades entre variáveis binárias são calculadas por métodos específicos, não sendo apropriadas as medidas utilizadas para as variáveis escaladas em intervalos.

## Medidas de Similaridade e Dissimilaridade

Para a formação de grupos de objetos onde os elementos de cada grupo possuam mais similaridades entre si do que em relação aos demais grupos, é necessário quantificar a similaridade entre os objetos.

As medidas de similaridade são valores numéricos que quantificam a distância entre dois objetos. Quanto menor o valor, mais semelhantes serão os objetos.

Não há uma medida de similaridade que sirva para todos os tipos de variáveis que podem existir numa base de dados. A seguir, temos as medidas de similaridade mais conhecidas e a associação para cada tipo de variável, segundo HAN et al. (2001).

### Distância Euclidiana

CARLANTONIO (2001) coloca que considerando os objetos  $x$  e  $y$  com apenas dois atributos ( $x \rightarrow x_i$  e  $x_j$ ;  $y \rightarrow y_i$  e  $y_j$ ), a distância Euclidiana é a distância em linha reta entre os dois pontos que representam os objetos, conforme podemos ver na Figura 14:

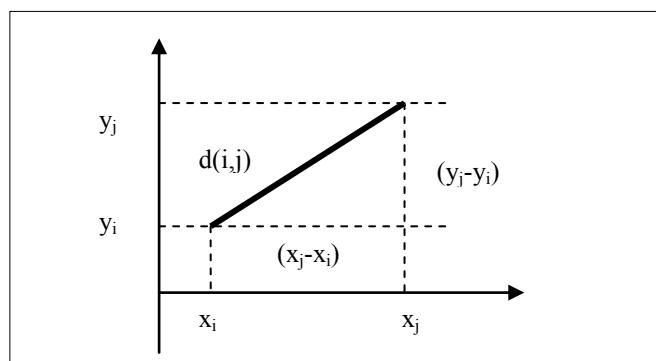


Figura 14 - Distância Euclidiana

Matematicamente, do teorema de Pitágoras, temos:

$$d(i, j) = \sqrt{(y_j - y_i)^2 + (x_j - x_i)^2}.$$

Generalizando para dois objetos com  $n$  atributos, temos:

$$d(i, j) = \sqrt{|x_{j1} - x_{i1}|^2 + |x_{j2} - x_{i2}|^2 + \dots + |x_{jn} - x_{in}|^2}.$$

Para que a unidade de medida dos atributos das variáveis não interfira na análise, é preciso normalizar os dados. A normalização faz com que todas as variáveis tenham o mesmo peso. Este passo fica a critério do usuário. Maiores informações sobre normalização podem ser encontradas em HAN et al. (2001) e CARLANTONIO (2001).

Segundo CARLANTONIO (2001), a distância Euclidiana normalmente é utilizada para medir a similaridade em variáveis escaladas em intervalos, mas também pode ser usada para medir a dissimilaridade em variáveis ordinais.

Em variáveis ordinais a diferença é que o valor de cada atributo é trocado pelo valor que ele ocupa na sequência dos dados possíveis. Exemplo: segunda-feira = 1, terça-feira = 2, quarta-feira = 3, ..., domingo = 7.

Logo em seguida, mapeamos cada valor de atributo em um novo valor contido na faixa [0,0; 1,0]. Isto é feito para que cada atributo tenha o mesmo peso. O mapeamento é feito através da seguinte expressão matemática:

$$z_{if} = \frac{r_{if} - 1}{m_f - 1},$$

onde  $i$  representa o objeto em questão e  $m_f$  é o número de possíveis estados para a variável  $f$ . Estes serão os valores normalizados dos atributos utilizados para calcular a dissimilaridade.

### Distância de Manhattan

Na distância de Manhattan, a distância entre dois pontos é a soma das diferenças absolutas de suas coordenadas. Seja os objetos  $x$  e  $y$  com apenas um atributo cada, ( $x \rightarrow x_i$  ;  $y \rightarrow y_i$ ), sua medida de similaridade é dada pela fórmula:

$$d(x, y) = \sum |x_i - y_i|, \text{ onde } x_i \text{ e } y_i \text{ são as coordenadas de cada objeto.}$$

A distância de Manhattan, assim como a Euclidiana, é utilizada em variáveis escaladas em intervalos e em variáveis ordinais.

Graficamente, podemos ver na Figura 15 a comparação das medidas de distância Euclidiana e de Manhattan.

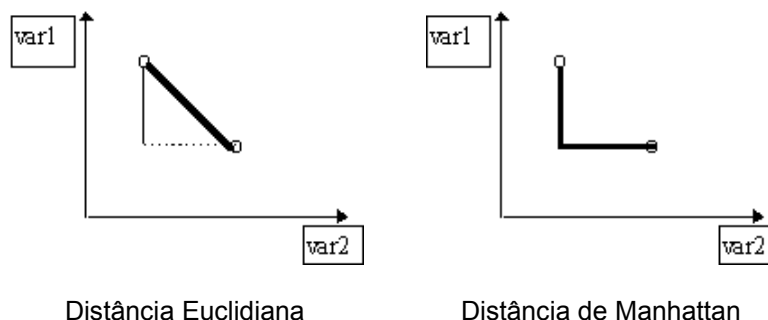


Figura 15 - Comparação de medidas de Similaridade

Segundo CARLANTONIO (2001), essa medida de similaridade é mais facilmente calculada do que a Euclidiana, mas ela pode não ser adequada se os atributos estão correlacionados, pois não há garantia da qualidade dos resultados obtidos.

## Matriz de Dissimilaridade

No caso de variáveis booleanas, ou binárias, onde os atributos assumem apenas dois valores distintos (1 ou 0, sim ou não, verdadeiro ou falso, presente ou ausente, etc.), tratar estes valores como numéricos pode levar a divergência em análises de agrupamentos.

Para calcular a distância  $d(i,j)$  de dois objetos (A e B), usamos uma matriz de dissimilaridade – cada elemento da matriz representa a distância entre pares de objetos, conforme mostra a Figura 16 – adaptada de CARLANTONIO (2001).

		Objeto "A"		soma
		presente (1)	ausente (0)	
Objeto "B"	presente (1)	a	b	a+b
	ausente (0)	c	d	c+d
soma		a+c	b+d	p

Figura 16 - Quantificação da similaridade/dissimilaridade entre os objetos "A" e "B"

De acordo com a tabela, temos:

- ✓  $a$  é o número de atributos presentes (valor = 1) para o objeto "A" e "B"
- ✓  $b$  é o número de atributos presentes (valor = 1) para o objeto "B" e ausente (0) para "A"
- ✓  $c$  é o número de atributos ausentes (valor = 0) para o objeto "B" e presente (1) para "A"
- ✓  $d$  é o número de atributos ausentes (valor = 0) para o objeto "A" e "B"
- ✓  $p$  é o número total de atributos ou caracteres comparados. Portanto  $p = a+b+c+d$ .
- ✓  $a+d$  é o número de coincidências
- ✓  $b+c$  é o número de divergências

Em CARLANTONIO (2001) vemos que há um grande número de coeficientes de similaridade e/ou dissimilaridade disponíveis na literatura. Nesta pesquisa, abordaremos os coeficientes de similaridade e os coeficientes de associação.

## Coefficientes de similaridade

Estes coeficientes ( $s$ ) baseiam-se na comparação das coincidências ( $a+d$ ) com o número total de atributos comparados ( $p$ ). O coeficiente de dissimilaridade é o valor complementar deste coeficiente de similaridade ( $1 - s$ ).

Os coeficientes de similaridades podem considerar ou não a ausência ( $d$ ) conjunta. Na Tabela 4 e Tabela 5, adaptadas de MEYER (2002), temos alguns dos coeficientes disponíveis na literatura.

<i>Coeficientes(s)</i>	<i>Fórmula</i>	<i>Intervalo de Ocorrência</i>
Simple Matching (1958)	$\frac{a+d}{p}$	[0,1]
Russel e Rao (1940)	$\frac{a}{p}$	[0,1]
Rogers e Tanimoto (1960) <sup>8</sup>	$\frac{a+d}{a+d+2(b+c)}$	[0,1]
Hamann (1961)	$\frac{(a+d)-(b+c)}{p}$	[-1,1]
Ochiai II (1957)	$\frac{ad}{\sqrt{(a+d)(a+c)(d+b)(d+c)}}$	[0,1]
Sokal e Sneath (1963) <sup>9</sup>	$\frac{2(a+d)}{2(a+d)+b+c}$	[0,1]

Tabela 4- Coeficientes de similaridade que consideram a ausência conjunta

Segundo CARLANTONIO (2001) e MEYER (2002), estes coeficientes possuem propriedades semelhantes por considerarem a ausência conjunta ( $d$ ). Suas variações se dão ao nível de importância dada tanto à ausência ( $d$ ) quanto à presença conjunta ( $a$ ) e divergências ( $b+c$ ). São utilizados principalmente quando os valores 0 ou 1 são igualmente importantes para a análise de clusters (atributos simétricos). Por exemplo, o atributo “sexo” é simétrico, pois os dois valores M e F são igualmente importantes.

<i>Coeficientes (s)</i>	<i>Fórmula</i>	<i>Intervalo de Ocorrência</i>
Jaccard (1908)	$\frac{a}{a+b+c}$	[0,1]
Anderberg (1973) <sup>10</sup>	$\frac{a}{a+2(b+c)}$	[0,1]

<sup>8</sup> Define peso duplo às situações discordantes, inclusão das ausências simultâneas.

<sup>9</sup> Define peso duplo às presenças e ausências simultâneas.

<sup>10</sup> Define peso duplo às situações discordantes, exclusão das ausências simultâneas.

Kulczynski I (1927) <sup>11</sup>	$\frac{a}{b+c}$	[0,+∞]
Kulczynski II (1927)	$\frac{a}{2} \left( \frac{1}{a+b} + \frac{1}{a+c} \right)$	[0,1]
Sorensen-Dice (1945) <sup>12</sup>	$\frac{2a}{2a+b+c}$	[0,1]
Ochiai (1957)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0,1]

Tabela 5 - Coeficientes de similaridade que desconsideram a ausência conjunta

Segundo Clifford & Stephenson (1975), citado em MEYER (2002), o uso de coeficientes que estão restritos ao intervalo [0,1] são mais adequados por não serem sensíveis a pequenas variações, principalmente em “a”.

### Coeficientes de associação

Estes coeficientes mostram como os pares de indivíduos estão associados e expressam a probabilidade de acontecimento, por acaso, de certo número de caracteres comuns a dois objetos. Variam de -1 a 1. Alguns destes coeficientes são descritos por MEYER (2002) e estão descritos na Tabela 6.

Coeficientes (s)	Fórmula	Coeficientes de Dissimilaridade
Yule	$\frac{ad - bc}{ad + bc}$	$\frac{\left(1 - \frac{ad - bc}{ad + bc}\right)}{2}$
Pearson	$\frac{ad - bc}{(a+b)(c+d)(a+c)(b+d)}$	$\frac{\left(1 - \frac{(ad - bc)}{(a+b)(c+d)(a+c)(b+d)}\right)}{2}$
McConnaughy	$\frac{(a^2 - bc)}{(a+b)(a+c)}$	$\frac{\left(1 - \frac{(a^2 - bc)}{(a+b)(a+c)}\right)}{2}$

Tabela 6 - Coeficientes de associação contidos no intervalo [-1,+1]

<sup>11</sup> Define o quociente entre presenças simultâneas e situações discordantes, exclusão das ausências simultâneas.

<sup>12</sup> Define peso duplo às presenças simultâneas, exclusão das ausências simultâneas.

Além de serem usados para variáveis booleanas, com algumas adaptações, os coeficientes de similaridade e associação também podem ser utilizados em variáveis nominais (categóricas).

Neste caso, consideramos a quantidade de atributos que possuem o mesmo valor no objeto “A” e no objeto “B”  $\rightarrow m$ , e a quantidade total de atributos  $\rightarrow p$ . Desta forma, calculamos a dissimilaridade entre dois objetos como:

$$d(A, B) = \frac{p - m}{p}, \text{ sendo possível utilizar pesos para aumentar o efeito de } m.$$

Segundo HAN et al. (2001), quando um algoritmo que trabalha com matrizes de dissimilaridade recebe uma matriz de dados, ele primeiro transforma-a em uma matriz de dissimilaridade antes de iniciar suas etapas de clusterização.

## Algoritmo: K-Means

O k-means é um algoritmo clássico de classificação não supervisionada. Segundo HAN et al. (2001) e BISHOP (2006), é muito popular devido sua facilidade de implementação e, apesar de sua eficiência, possui a limitação de trabalhar somente com valores numéricos.

Neste algoritmo é necessário informar o número de clusters em que os dados deverão ser categorizados. A estimativa dos centros dos clusters afeta não apenas na convergência do algoritmo, mas também na precisão dos centróides obtidos no final.

Normalmente, quando se conhece a natureza dos dados, a escolha dos centros dos grupos não é feita de forma aleatória, aproveita-se o conhecimento prévio dos dados.

O funcionamento dele é descrito resumidamente por particionar os objetos em k clusters e a partir da similaridade do valor da média dos atributos numéricos, agrupa os demais objetos da base de dados nestes clusters previamente indicados, HAN et al. (2001) .

Em HAN et al. (2001) temos que o algoritmo pode ser resumindo em cinco passos:

1. Selecione  $C$  pontos como centróides iniciais ( $c_k$ )
2. Repeat
3. Forme  $k$  clusters associando cada objeto a seu centróide mais próximo a partir da função objetivo  $J$ .
4. Recalcule o centróide de cada cluster
5. Until Centróides não apresentam mudanças

Considerando o algoritmo acima, no primeiro passo é escolhido aleatoriamente  $C_k$  objetos como centróides para o agrupamento. A Figura 17 ilustra o processo e representa um conjunto de dados com  $k = 3$ .

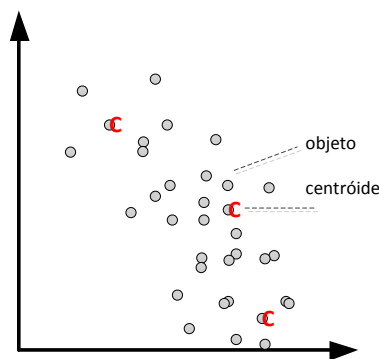


Figura 17 - Definição do número de centróides.

Em seguida, agruparemos todos os objetos através do cálculo distância dos mesmos até cada centróide, utilizando a função objetivo ( $J$ ), conforme visto na Figura 18. A função objetivo  $J$  do modelo pode ser escrita da seguinte forma:

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x(n) - c_k\|^2$$

Nela, um determinado elemento  $\mathbf{x}(n)$  pertence ao grupo  $\omega_k$  se sua distância ao centróide  $\mathbf{c}_k$  for menor que ao centróide  $\mathbf{c}_j$ . A variável  $r_{nk}$  pode assumir valores discretos 0 ou 1:

$$r_{nk} = 1 \text{ se } x(n) \in \omega_k \text{ ou } r_{nk} = 0 \text{ se } x(n) \notin \omega_k$$

Desta forma, é possível perceber que no k-means um elemento pode pertencer somente a um grupo.

De acordo com HAN et al. (2001), o objetivo é minimizar a função objetivo  $J$  sabendo que as duas variáveis que podemos ajustar no modelo são:  $r_{nk}$  e  $\mathbf{c}_k$ .

Com os valores estimados de  $\mathbf{c}_k$ , minimizaremos a função  $J$  deixando  $\mathbf{c}_k$  fixo e então encontraremos os valores de  $r_{nk}$ .

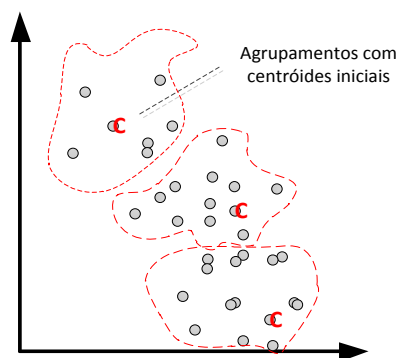


Figura 18 - Agrupamento com os centróides iniciais.

Concluído o primeiro agrupamento, fazemos o contrário da etapa anterior: mantemos  $r_{nk}$  fixo e encontramos os novos valores de  $\mathbf{c}_k$  pela média aritmética das distâncias de todos os elementos do grupo até seu respectivo centróide – Figura 19:

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

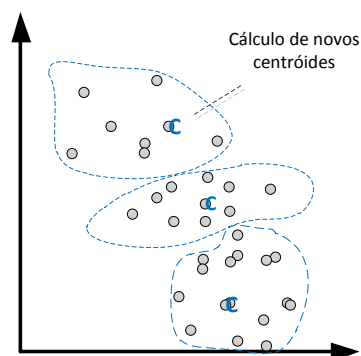


Figura 19 - Cálculo de novos centróides.

Depois de encontrado os novos centróides, um novo agrupamento é encontrado a partir da função objetivo ( $J$ ).

Na Figura 20 temos a variação dos objetos em relação aos seus respectivos centróides.

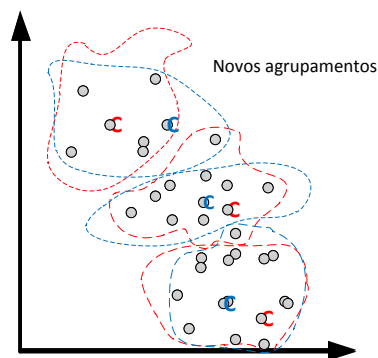


Figura 20 - Agrupamento com os novos centróides.

O processo deve repetir até que não haja mais mudanças, ou seja, até que a média aritmética das distâncias de todos os elementos do grupo até seu respectivo centróide não altere a posição do centróide.

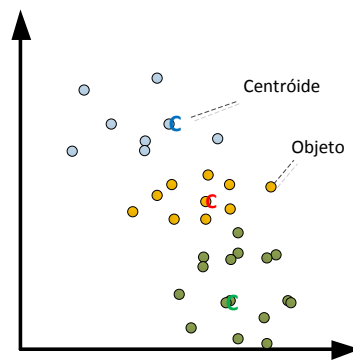


Figura 21 - Convergência do algoritmo de clusterização.

Podem acontecer convergências prematuras através de mínimos locais, por isso, é importante rodar o algoritmo diversas vezes e comparar os resultados.

O algoritmo k-means está disponível no software SPSS (Statistical Package for the Social Sciences), e tem como particularidade prática o tratamento feito para dados faltantes nas variáveis do tipo “range” ou “flag”. Para ambos os casos, os campos (em branco ou nulo) são substituídos por “0,5”.

Detalharemos a seguir os parâmetros básicos do modelo que podem ser configurados pelo usuário, segundo o manual do usuário da ferramenta – Figura 22:

- ✓ *Model name*: especifica o nome do modelo que será produzido. Caso o usuário deixe selecionada a opção “auto”, o SPSS usará o nome padrão “Kmeans”.
- ✓ *Specified number of clusters*: Neste campo o usuário deve especificar o número de agrupamentos. O valor padrão do algoritmo são cinco agrupamentos.
- ✓ *Generate distance field*: Caso esta opção esteja selecionada, o modelo incluirá um campo com a distância de cada registro até o centro de cada cluster.

- ✓ *Show cluster proximity*: Esta opção permite que sejam informadas no modelo as distâncias entre os centros de cada cluster.

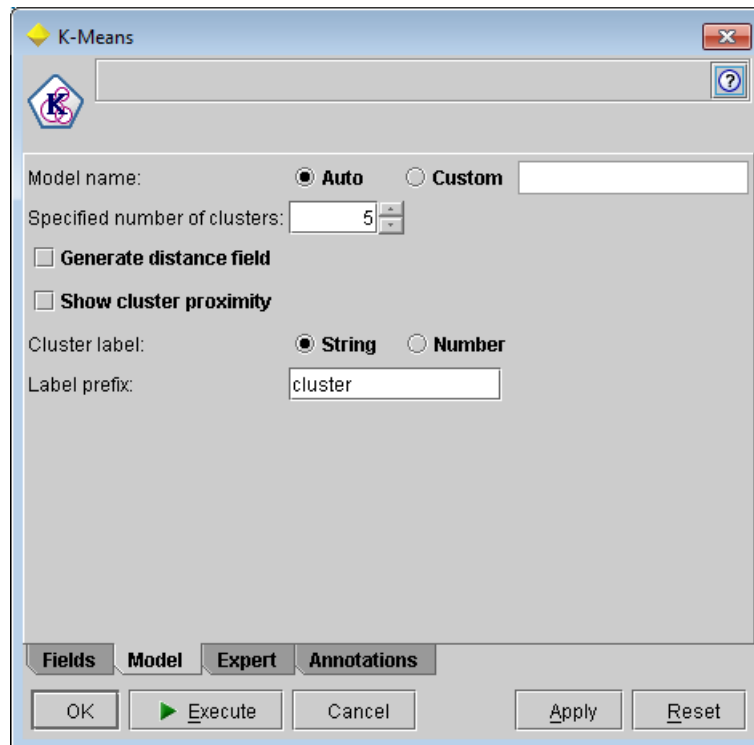


Figura 22 - Parâmetros básicos de configuração do k-means no software SPSS.

- ✓ *Cluster label*: Especifica a nomenclatura dos grupos criados. Os grupos podem ser identificados por uma string “cluster”, por exemplo. Assim, cada agrupamento criado será identificado como "cluster 1", "cluster 2", etc. Ou podem ser identificados simplesmente através de números, caso seja selecionado “Number”.

Além dos parâmetros básicos, é possível acessar a aba de configurações avançadas (“Expert”), Figura 23, que possui as seguintes opções de ajuste do modelo:

- ✓ *Stop on*: Especifica os critérios de parada do algoritmo. Por padrão, a convergência se dá após 20 iterações ou quando a variação da distância dos centroides for menor que 0.000001, o que ocorrer primeiro. Para alterar os parâmetros dos critérios de parada, basta selecionar a opção “custom”.
  - *Maximum Iterations*: representa o número máximo de iterações do algoritmo. Se ou quando este limite for atingido o algoritmo terminará e exibirá o conjunto de clusters obtido.
  - *Change tolerance*: define o valor máximo de variação das medidas de distância do centro de cada cluster, em cada iteração.

Esta regra é avaliada da seguinte forma:  $\max_j \|C_j(t) - C_j(t - 1)\|$ , onde  $C_j(t)$  é o vetor de centroides na iteração  $t$  e  $C_j(t - 1)$  o vetor da iteração anterior. Se a maior mudança nesta iteração for menor que o especificado neste campo o algoritmo terminará e exibirá o conjunto de clusters obtido.

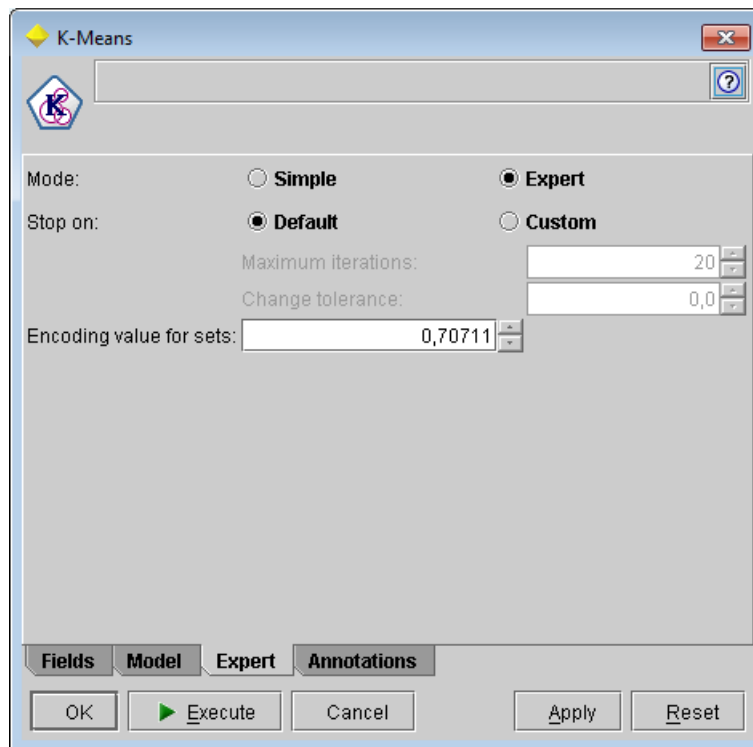


Figura 23 - Parâmetros avançados de configuração do k-means no software SPSS.

- ✓ *Encoding value for sets*: Este campo controla o peso que as variáveis do tipo “set” possuem no algoritmo k-means. O tipo SET é usado para representar uma coleção de variáveis cujo número é conhecido e modesto e cuja ordem não é significativa. Por padrão, o valor é:  $\sqrt{0,5}$  (aproximadamente 0,707), e fornece um balanceamento entre as variáveis do tipo “range” (intervalo de variação) e “set”.
  - Para dar maior peso as variáveis do tipo “set”, o valor deste campo deve ser próximo a 1.
  - Para dar maior peso as variáveis do tipo “range”, o valor deste campo deve ser próximo a 0.

Segundo CARLANTONIO (2001), quando há ocorrência de variáveis categóricas, algumas aproximações são usuais: transformá-las em contínuas, atribuindo valores numéricos às suas categorias, ou em binárias, fazendo com que cada uma das suas categorias se torne uma variável que represente presença ou ausência desse determinado atributo, transformar as contínuas em categóricas criando classes de valores ou ainda aplicar aos dados medidas específicas que tratam as observações conjuntamente.

## Algoritmo: Dois Estágios – Two Step Cluster

O método de agrupamento de dois estágios, também conhecido por two-step cluster analysis, está disponível no software estatístico SPSS, utiliza o mesmo princípio do algoritmo k-means, que busca formar grupos de registros homogêneos com o máximo de heterogeneidade entre grupos possível, mas se destaca por permitir agrupar dados sem que seja necessário informar ao algoritmo o número de classes.

O algoritmo TwoStep é concluído em duas etapas distintas. Na primeira etapa, faz uma leitura única de todas as variáveis e associa cada indivíduo a um cluster. Em seguida, na segunda etapa, os clusters são progressivamente mesclados por um método de agrupamento hierárquico aglomerativo, que cria grupos cada vez maiores, de forma a não haver a necessidade de termos outra passagem através os dados.

A vantagem de utilizar o agrupamento hierárquico na segunda etapa é que simplifica o processo. Os métodos hierárquicos aglomerativos, conforme explicado anteriormente, começam com um número de clusters igual ao número de indivíduos e, a partir da distância entre eles, faz o agrupamento recursivo até que se tenha um único grupo que represente toda a massa de dados.

Este tipo de abordagem muitas vezes não apresenta bom desempenho com grandes quantidades de dados. Além disso, o modelo é influenciado pela ordem na qual a informação é lida no banco de dados. Reordenar os dados e reconstruir o modelo pode levar a resultados diferentes de agrupamentos.

Outra questão importante refere-se aos requisitos do modelo. O agrupamento de dois estágios não lida com valores ausentes. Todos os atributos em branco de qualquer variável de entrada serão ignorados quando a construção do modelo.

Dentre os pontos fortes deste método de agrupamento, destacamos:

- ✓ O algoritmo TwoStep Cluster é capaz de lidar com diversos tipos de variáveis;
- ✓ É robusto, consegue manipular grandes conjuntos de dados de forma eficiente;
- ✓ Tem a capacidade de testar várias soluções de cluster e escolher o melhor, com isso o usuário não precisa determinar a quantidade de grupos (k) inicial.
- ✓ Pode ser configurado para excluir automaticamente *outliers*, que, devido à característica do algoritmo, pode interferir nos resultados.

Detalharemos a seguir os parâmetros do clusterizador que podem ser ajustados, segundo o manual do usuário, no Clementine SPSS – Figura 24:

- ✓ *Model name*: especifica o nome do modelo que será produzido. Caso o usuário deixe selecionada a opção “auto”, o SPSS usará o nome padrão “TwoStep”.
- ✓ *Standardize numeric fields*: Por padrão, o algoritmo normalizará todas as variáveis numéricas da base de dados para o intervalo [0,1]. Para manter a escala original dos dados é necessário desmarcar esta opção.

- ✓ *Exclude outliers*: outlier (ou anomalia) é uma observação no conjunto de dados que é considerada dissimilar ou aberrante em relação ao restante dos dados. É um valor normalmente muito grande ou muito pequeno quando comparado com os demais do mesmo conjunto, e pode ter sido resultado de um erro de medida ou, então, pode ser um indicativo de um comportamento atípico dos dados sob determinadas condições.

A detecção de outlier é feita na etapa de pré-agrupamento. Quando essa opção é selecionada, subgrupos com poucos registros em relação a todos os outros subgrupos são considerados outliers em potencial e a árvore de agrupamentos é refeita excluindo estes registros. Alguns destes subgrupos de outliers em potencial podem ser agrupados a cada etapa de recursividade do algoritmo se forem similares o bastante entre si. O total de subgrupos de outliers que não foram aglutinados é adicionado a um cluster específico chamado "noise" (ruído em inglês) e excluídos da etapa de agrupamento hierárquico.

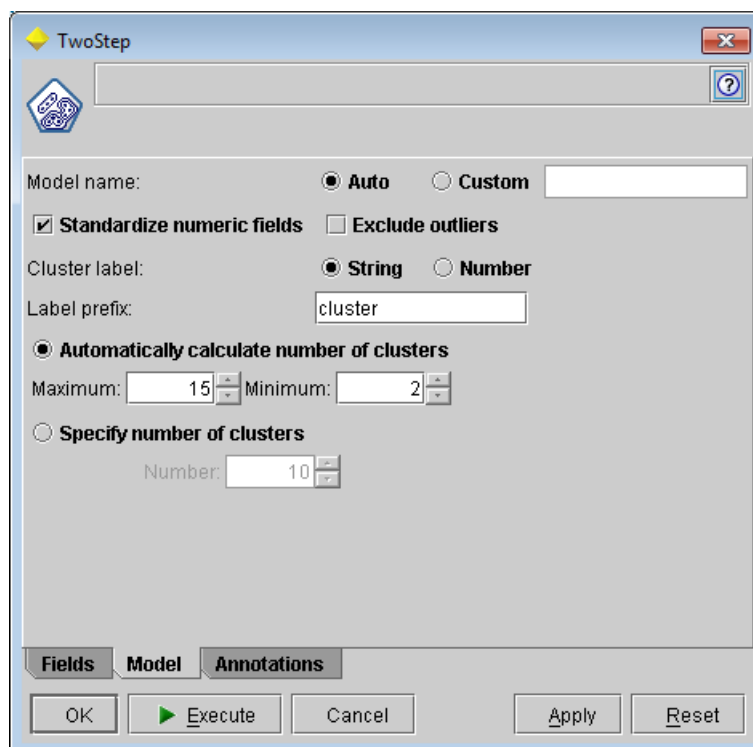


Figura 24 - Parâmetros do TwoStep Cluster - SPSS

- ✓ *Cluster label*: Especifica a nomenclatura dos grupos criados. Os grupos podem ser identificados por uma string “cluster”, por exemplo. Assim, cada agrupamento criado será identificado como "cluster 1", "cluster 2", etc. Ou podem ser identificados simplesmente através de números, caso seja selecionado “Number”.
- ✓ *Automatically calculate number of clusters*: O algoritmo TwoStep Cluster permite analisar grande quantidade de soluções de agrupamento a partir da estimativa de um número máximo e mínimo de clusters e retorna o melhor número de agrupamentos para o conjunto de dados processados. O usuário deve, neste momento, estimar o range de clusters em que os dados serão submetidos.

- ✓ *Specify number of clusters*: Esta opção é utilizada apenas quando o número de clusters é conhecido a priori. Na verdade, utilizar o algoritmo de dois estágios com esta opção selecionada não faz muito sentido, mas é um recurso disponível no software SPSS.

O software SPSS disponibiliza três diferentes métodos de agrupamento: análise de grupos hierárquicos, k-means e two-step cluster.

O algoritmo two-step é recomendado em base de dados grandes, com mais de 1.000 registros e vários tipos de variáveis (numéricas, ordinais, categóricas, etc.). Para bases de dados pequenas, onde o objetivo é explorar soluções diferentes de agrupamentos, o melhor algoritmo é agrupamento hierárquico. E, quando se tem uma idéia do número de agrupamentos desejado, além de uma base de dados de tamanho moderado, recomenda-se o uso do k-means.

### Índices de Validação de Clusters

A validação de agrupamentos tem como objetivo avaliar qual cluster possui uma melhor estrutura de agrupamento em um conjunto de dados não conhecido. Abaixo temos a descrição dos índices de avaliação mais comuns.

- **Índice PBM**

O índice de validação de agrupamento de dados conhecido por PBM, cujo nome é formado pelas iniciais dos sobrenomes dos autores Pakhira, Bandyopadhyay e Maulik em OLIVEIRA (2002).

É definido nas equações como:

$$PBM(K) = \left( \frac{1}{K} \times \frac{E_i}{E_K} \times D_K \right)^2$$

$$E_k = \sum_{i=1}^k E_i,$$

$$E_k = \sum_{i=1}^n u_{ii} d(x_i, w_i),$$

$$D_k = \max_{i,j=1}^k d(w_i, w_j).$$

Onde K é o número de clusters, n é o número de registros em um conjunto de dados e  $w_i$  é o centro do  $i$ -ésimo agrupamento. Ou seja, pondera o número de clusters (K), a relação entre a densidade dos dados e a densidade dos clusters e a distância máxima entre os centros das classes.

O objetivo é maximizar este índice a fim de se obter o número real de agrupamentos, ou seja, o valor máximo deste índice indica o melhor particionamento, OLIVEIRA (2002).

- **Índice Calinski-Harabasz**

O índice Calinski-Harabasz (CH), também conhecido como índice VRC (Variance Ratio Criterion) faz a relação entre a soma dos quadrados das distâncias dos centróides entre clusters  $B(K)$  e dentro dos clusters  $W(K)$ .

Para melhor compreender a definição do índice CH, CALINSKI et al. (1974), seja  $X = \{x_1, \dots, x_n\}$  o conjunto dos vetores de atributos dos objetos de uma base de dados e  $W$  definido como

$$W = \sum_{i=1}^k \sum_{x \in C_i} \|x - \bar{x}_i\|_E^2,$$

em que  $\bar{x}_i$  é o centróide do grupo  $C_i$  (média aritmética dos vetores de medidas que descrevem os objetos do grupo  $C_i$ ) e  $\|\cdot\|_E$  é a norma Euclidiana. Note que, quanto mais compactos os grupos, menor o valor de  $W$ . Seja  $B$  definido como

$$B = \sum_{i=1}^k |C_i| \|\bar{x}_i - \bar{x}\|_E^2,$$

em que  $\bar{x}$  é o centróide da base de dados (média aritmética dos valores de  $X$ ) e  $|C_i|$  é a quantidade de objetos do grupo  $C_i$ . Note que o valor de  $B$  tende a aumentar com a distância entre os grupos.

Finalmente, o índice CH (ou VRC) é definido como  $CH = VRC = \frac{B/(k-1)}{W/(N-k)}$ ,

onde o objetivo é maximizar o índice a fim de se obter classes densas e separadas – CALINSKI et al. (1974).

É importante ressaltar que o índice CH, como originalmente formulado, não é adequado para o contexto das bases de dados relacionais, uma vez que lida com os vetores de medidas  $X$ .

## 2.4.2) Regras de associação

As regras de associação, conforme visto em AGRAWAL et al. (1993), são padrões descritivos eficazes na busca por relações entre conjuntos de itens em uma transação. Uma regra é representada por uma expressão do tipo  $X \rightarrow Y$ , onde  $X$  e  $Y$  são conjuntos de itens, ou seja:

$$X_1, X_2, \dots, X_m \rightarrow Y_1, Y_2, \dots, Y_n.$$

Intuitivamente, as regras significam que as transações da base de dados que contêm  $X$  tendem a conter  $Y$ .

Segundo AGRAWAL et al. (1993) e AGRAWAL et al. (1994), as aplicações típicas de análise de regras de associações incluem:

- ✓ Análise de cestas de mercado, onde  $X$  e  $Y$  representam produtos de uma “cesta” e o objetivo da análise é identificar combinações de produtos comprados frequentemente.
- ✓ Auxílio na previsão de complicações com certos tipos de tratamentos médicos.
- ✓ Análise de combinações incomuns de resgate de seguros, que poderiam indicar o sinal de fraude.

Um modelo formal para representar o problema de descoberta de regras de associação considera um conjunto de atributos binários  $I = \{I_1, I_2, \dots, I_m\}$  chamado de itens, e, um conjunto  $D$  de transações, onde cada transação  $T$  é um conjunto de itens tal que  $T \subseteq I$ .

Associado com cada transação está um atributo que a identifica unicamente, chamado  $T_{ID}$ . Uma transação  $T$  contém  $X$ , sendo  $X$  um conjunto de itens em  $I$ , se  $X \subseteq T$ .

Uma regra de associação é uma implicação do tipo:

$$X \rightarrow Y, \text{ onde } X \subset I, Y \subset I \text{ e } X \cap Y = \emptyset.$$

A regra  $X \rightarrow Y$  tem suporte  $s$  em  $D$ , se  $s\%$  das transações em  $D$  contêm  $X$  e  $Y$ . O suporte é a probabilidade de uma transação em  $D$  conter  $X \cup Y$ , ou seja, ele indica a frequência da regra. O suporte  $s(X \cup Y)$  é dado, então, por AGRAWAL et al. (1993):

$$\text{suporte} = \text{quantidade} \frac{(X \cup Y)}{\text{tamanho}(D)} = \frac{|\{t \in D | X \subset t, Y \subset t\}|}{|D|}$$

A regra  $X \rightarrow Y$  é válida no conjunto de transações  $D$  com o grau de confiança  $c$ , se  $c\%$  das transações em  $D$  que contêm  $X$  também contêm  $Y$ . A confiança é a probabilidade de  $Y$  ocorrer em uma transação de  $D$  visto que  $X$  ocorre, ou seja, indica a “força” da regra.

Ainda em AGRAWAL et al. (1994), dado um conjunto de transações, o problema na mineração por regras de associação está em gerar todas as regras que sejam realmente

úteis. Regras úteis são regras que ocorrem com frequência, são confiáveis e fazem previsões interessantes.

Regras que refletem casos que raramente ocorrem tendem a ser desprezadas. Por este motivo, deve-se estipular um valor mínimo para o suporte, geralmente referido como *suporte\_mínimo*. O suporte de uma regra  $X \rightarrow Y$ , dado por  $\text{suporte}(X \cup Y)$ , é a probabilidade de uma transação satisfaça tanto X como Y. Caso o suporte de um conjunto de itens for maior ou igual ao mínimo estabelecido ( $\text{suporte}(X) \geq \text{suporte\_mínimo}$ ), diz-se que este conjunto de itens é freqüente. Um exemplo de regra eficiente seria:

*75% das pessoas que compram Coca-Cola compram batata-frita.*

*60% das pessoas compram batata-frita e sorvete.*

Por outro lado, as regras também precisam refletir a realidade com certo grau de certeza de estarem certas. Regras que raramente estão corretas, ou seja, que a implicação raramente se confirma, também não são regras úteis. Portanto, deve-se estipular um valor mínimo para a confiança da regra, geralmente referido como *confiança\_mínima*.

AGRAWAL et al. (1993) coloca que a confiança é dada pela fórmula abaixo e representa a probabilidade de que uma transação satisfaça Y, dado que ela satisfaz X:

$$\text{confiança} = \frac{\text{suporte}(X \cup Y)}{\text{suporte}(X)}$$

Segundo ZAKI et al. (1998), a tarefa de descobrir regras de associação consiste em extrair do banco de dados todas as regras com valor de Suporte e Confidência maiores ou iguais ao especificado pelo usuário, e pode ser decomposta em dois passos:

No primeiro passo, o algoritmo determina todos os conjuntos de itens que têm um suporte acima do suporte mínimo estabelecido pelo usuário. Estes conjuntos são denominados de conjuntos de itens freqüentes. É a etapa que mais consome recursos de processamento.

No segundo passo, gera as regras de associação para cada conjunto de itens freqüentes, selecionando apenas as regras que possuam o grau de confiança mínimo estipulado na etapa anterior.

Na geração das regras, a base de dados não necessita ser percorrida, pois o cálculo da confiança da regra pode ser feito somente com o suporte do antecedente e do conseqüente da regra sendo analisada. Na fase de geração dos itens freqüentes, o suporte de todos os itens já é calculado.

Um dos problemas que se pode enfrentar ao aplicar a associação, conforme visto em AGRAWAL et al. (1993), AGRAWAL et al. (1994) e ZAKI et al. (1998), é a geração de um grande número de regras não interessantes ou redundantes, juntamente com as regras úteis, até porque, quanto maior for a quantidade de itens na base de dados, maior será o número de combinações de regras possíveis. Uma forma de tratamento seria a

unificação de produtos similares, com a criação de grupos de produtos que apresentam características semelhantes.

Outro problema é o desempenho do algoritmo, pois a carga de processamento é grande, devido à grande quantidade de análises que são realizadas para encontrar os mais diversos tipos de combinações de associação, por esta razão, é necessária a aplicação de técnicas e algoritmos eficientes.

Nesta pesquisa utilizamos dois algoritmos de extração de regras de associação disponíveis no software SPSS: "Apriori Node" e "GRI Node". Detalharemos a seguir as particularidades de cada um, bem como seus parâmetros de ajuste do algoritmo.

### Algoritmo: Apriori Node

O método a priori extrai regras de associação que são declaradas da seguinte forma: Se antecedente(s) Então conseqüente(s). Por exemplo:

*"se um cliente compra um aparelho de barbear e loção pós-barba, então este cliente vai comprar creme de barbear com confiabilidade de 80%."*

Para criar um conjunto de regras a priori é necessário que o usuário especifique quais serão as variáveis que deverão ser analisadas na entrada, como antecedentes, e saída, como conseqüentes, conforme visto na Figura 25.

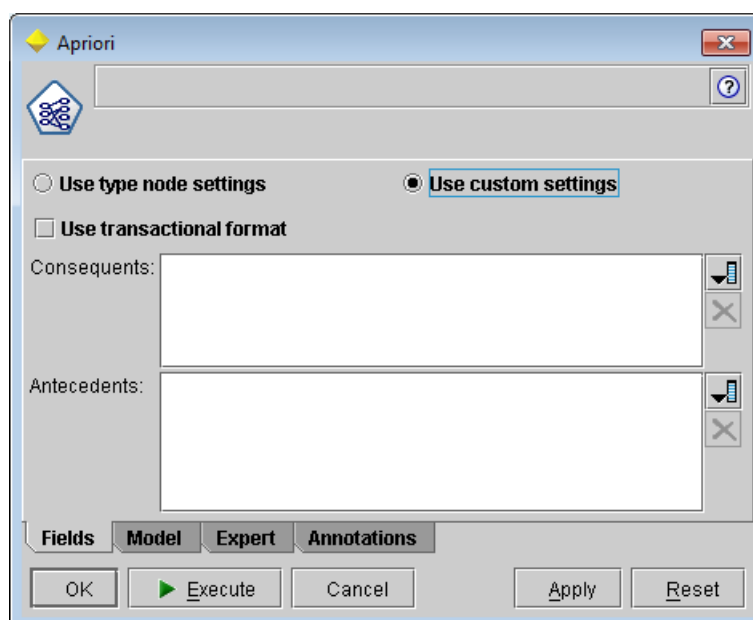


Figura 25 - Apriori node - SPSS

De acordo com o manual do SPSS, o software é otimizado para garantir a eficiência no processamento de grande quantidade de dados, principalmente se utilizado "Apriori Node", que, nesses casos, demonstra melhor desempenho que o outro método utilizado ("GRI Node").

Os parâmetros de ajustes básicos do modelo que podem ser configurados pelo usuário – Figura 26 - serão detalhados a seguir:

- ✓ *Model name*: especifica o nome do modelo que será produzido. Caso o usuário deixe selecionada a opção “auto”, o SPSS usará o nome padrão “Apriori”.
- ✓ *Minimum rule support*: este campo é usado para informar o valor suporte\_mínimo, usado como critério de escolha de regras úteis no conjunto de regras ao final da execução do algoritmo. O suporte refere-se ao percentual de registros no conjunto treinamento em que os antecedentes (a parte SE da regra) são verdadeiros.

Caso o algoritmo retorne um conjunto de regras aplicáveis a uma pequena parte do conjunto de dados, pode-se tentar aumentar este percentual.

Importante ressaltar que esta definição de suporte não se aplica a todos os algoritmos de extração de regras de associação do SPSS.

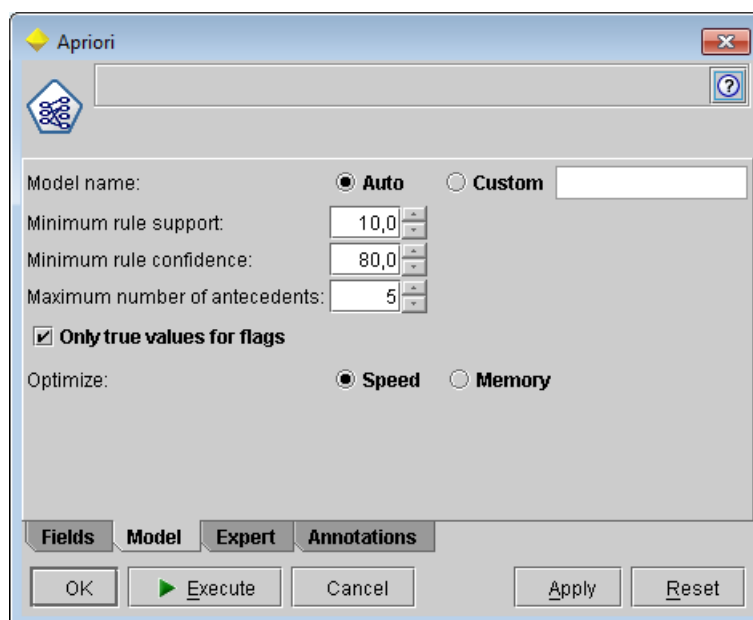


Figura 26 - Parâmetros de ajuste - Apriori node – SPSS

- ✓ *Minimum rule confidence*: este campo é usado para informar o valor confiança\_mínima. Conforme dito anteriormente, a confiança é a probabilidade de uma consequência ocorrer em função de uma premissa (antecedente) verdadeira, ou seja, indica a “força” da regra.

As regras com o valor de confiança menor que o critério especificado são descartadas. Este campo é muito útil quando o algoritmo retorna um grande número de regras, pois neste caso, é possível aumentar esta configuração. Da mesma forma, caso retorne poucas regras, este parâmetro pode ser reduzido.

- ✓ *Maximum number of antecedents*: é possível especificar o número máximo de condições prévias para qualquer regra, o que pode ser usado para limitar a

complexidade das mesmas. Se as regras se mostrarem muito complexas, este valor deve ser minimizado.

O campo "*Only true values for flags*" é utilizado quando a base de dados possui variáveis binárias, neste caso, usaria somente os valores verdadeiros (1), o que melhoraria o entendimento das regras.

No caso de "*Optimize*", o usuário pode decidir que recurso computacional prefere preservar: processador ou memória. Caso os experimentos desta pesquisa são feitos no ambiente de desenvolvimento, optou-se por otimizar o tempo de resposta ("*speed*").

Para aqueles que possuem grande conhecimento do algoritmo a priori, que pode ser visto em AGRAWAL et al. (1993), AGRAWAL et al. (1994) e ZAKI et al. (1998), é possível experimentar parâmetros avançados de ajustes do modelo. Estes parâmetros avançados ("*Expert*") alteram a forma na qual o algoritmo faz a avaliação do potencial de cada regra.

O SPSS disponibiliza cinco métodos diferentes de avaliação de regras. São eles: Rule Confidence, Confidence Difference, Confidence Ratio, Information Difference e Normalized Chi-square.

Abaixo, detalhamos os três mais utilizados durante a análise exploratória dos dados:

- 1) *Rule Confidence*: utiliza a acuracidade da regra como parâmetro de avaliação. Esta é a medida de avaliação padrão do modelo. Como o valor de "*Minimum rule confidence*" é definido das opções gerais do modelo, não há necessidade de especificar uma medida mínima para este critério, pois seria redundante.
- 2) *Confidence Difference*: esta medida de avaliação é a diferença absoluta entre a confiança da regra e sua confiança prévia. Esta opção minimiza o BIAS<sup>13</sup> nos casos onde as variáveis de consequência não são uniformemente distribuídas, e ajuda na eliminação de regras óbvias.

Caso esta medida de avaliação seja utilizada, será preciso estipular o valor de "*evaluation measure lower bound*", que representa o valor mínimo de confiança que será usado na avaliação da regra.

- 3) *Confidence Ratio*: esta medida de avaliação pode ser expressa como sendo igual a  $\frac{\text{confiança (regra)}}{\text{confiança}} - 1$ . Como na medida de avaliação anterior, este método leva em consideração as distribuições desiguais. Este critério é muito útil para encontrar regras que predizem eventos raros.

Em seguida temos a comparação do algoritmo Apriori com o GRI Node.

---

<sup>13</sup> Maiores detalhes sobre BIAS e Variância disponível em [1], BISHOP (2006) e HAN et al. (2001).

## Algoritmo: GRI Node

O algoritmo GRI (Generalized Rule Induction), assim como o anterior, também tem como objetivo extrair regras de associação em bases de dados no formato: Se antecedente(s) Então conseqüente(s).

O grande destaque para este método, em comparação ao anterior, é que o GRI pode lidar com mais de uma variável para cada conjunto de análise (entrada – antecedência, e saída - conseqüência). Além disso, o GRI pode lidar com variáveis numéricas como se fossem campos simbólicos.

O conjunto de regras obtidas geralmente é de fácil entendimento, e em alguns casos se sobrepõem, fazendo com que alguns registros desencadeiem mais de uma regra. Isso faz com que as regras sejam mais genéricas, possibilitando assim, o uso em conjunto de árvores de decisão.

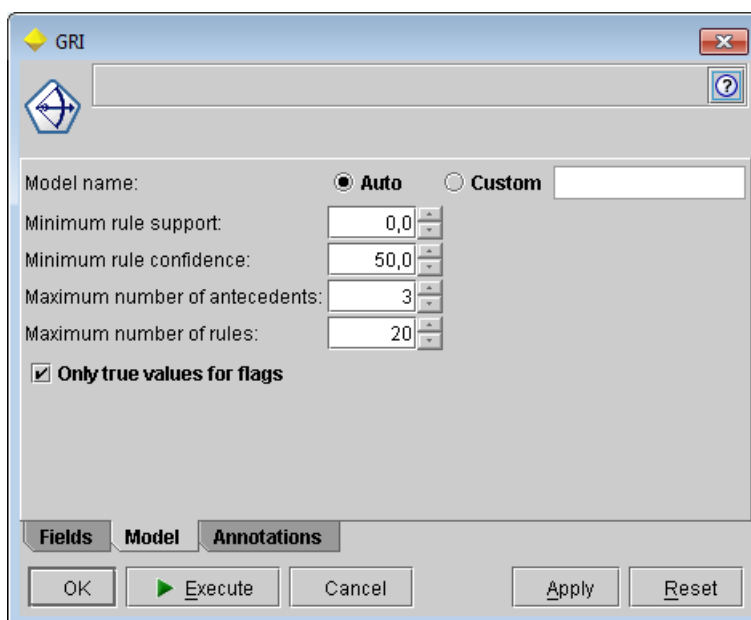


Figura 27 - GRI Node - SPSS

Os parâmetros de ajustes do modelo são praticamente os mesmos que o Apriori Node, conforme podemos identificar na Figura 27 – acima. O parâmetro que temos a mais neste algoritmo é:

- ✓ *Maximum number of rules*: esta opção define o número máximo de regras. As regras são exibidas em ordem decrescente de interesse (calculado pelo próprio algoritmo). Pode acontecer do algoritmo retornar um conjunto de regras menor que o especificado neste critério se os valores de confiança e suporte forem muito rigorosos.

Após a execução, o algoritmo retorna uma matriz que identifica cada transação e relaciona as conseqüências e antecedências, juntamente com o respectivo valor de suporte e confiança, conforme pode ser visto no exemplo da Tabela 7.

<b>Consequent</b>	<b>Antecedent</b>	<b>Support %</b>	<b>Confidence %</b>
Aprovador = EMCE360	prazoemissão do pedido > 172.0 and prazoemissão do pedido < 182.0	0.1	100.0
Aprovador = EMCE615	prazoemissão do pedido < -62.5 and prazoemissão do pedido > -161.0	0.03	100.0
Requisitante = EMCE373	prazo de chegada material < -73.0 and prazo de aprovação compra < 0.5	0.03	100.0
Requisitante = EMCE1132	prazoemissão do pedido > 220.5	0.03	100.0
UnidNegócio = 900	prazo de chegada material < -0.5 and prazo de liberação do material < 21.5	6.98	73.56
UnidNegócio = 900	prazo de chegada material < -0.5 and prazo de chegada material > -19.5	7.45	72.97
UnidNegócio = 900	prazo de chegada material < -0.5 and prazo de aprovação compra < 2.5 and prazo de chegada material > -18.5	5.94	72.88
UnidNegócio = 900	prazo de chegada material < -0.5 and prazoemissão do pedido > 1.5	6.61	71.57
UnidNegócio = 720	prazoemissão do pedido < -11.0 and prazo de aprovação compra < 0.5	0.1	66.67
UnidNegócio = 900	prazo de aprovação compra > 0.5 and prazoemissão do pedido > 1.5 and prazo de chegada material < 7.5	25.31	66.31
UnidNegócio = 900	prazo de aprovação compra > 0.5 and prazo de chegada material < 8.5	33.74	63.08
UnidNegócio = 900	prazo de aprovação compra > 0.5 and prazoemissão do pedido > 1.5	35.52	62.38
UnidNegócio = 900	prazo de aprovação compra > 0.5 and prazo de aprovação compra < 9.5	36.12	61.52
UnidNegócio = 900	prazo de aprovação compra > 0.5	45.55	60.65
Aprovador = EMCE360	prazoemissão do pedido > 172.0	0.23	57.14
UnidNegócio = 900	prazoemissão do pedido < -11.0	0.13	50.0
Aprovador = EMCE615	prazoemissão do pedido < -62.5	0.07	50.0
Requisitante = EMCE373	prazo de chegada material < -73.0	0.07	50.0

Tabela 7 - Exemplo de Regras de Associação – GRI Node

### 2.4.3) Classificação por Indução de Árvores de Decisão

Segundo STUART (2003), a árvore de decisão (do inglês decision tree) é uma maneira gráfica de visualizar o conjunto de condições necessárias para um determinado fim. Apresenta-se no formato de uma árvore, conforme visto na Figura 28, onde os nós quadrados representam decisões (fim), e os nós redondos, nós de incerteza, representam eventos aleatórios (condições).

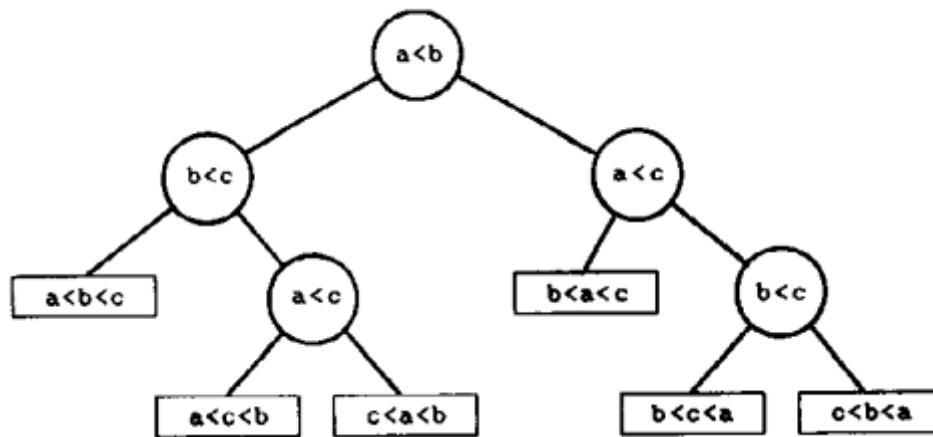


Figura 28 - Exemplo de árvore de decisão.

Generalizando o conceito, cada nó de decisão contém um teste num atributo. Cada ramo descendente corresponde a um possível valor deste atributo. Cada folha está associada a uma classe e cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

As regras são do tipo SE x, Então y e são muito usadas na implementação de sistemas especialistas.

Para que uma árvore seja construída é necessária uma grande sequência de testes. Cada nó da árvore corresponde a um teste do valor de uma das propriedades, e os ramos deste nó são identificados com os possíveis valores do teste. Cada nó folha da árvore especifica o valor de retorno se a folha for atingida.

A tarefa de construção de uma árvore de decisão é chamada de indução. Ainda segundo STUART (2003), a maioria dos algoritmos de indução de árvores de decisão constrói a árvore recursivamente de cima para baixo, ou seja, do nó raiz em direção aos nós terminais. Em cada iteração os algoritmos procuram pelo atributo que melhor separa as classes para realizarem a ramificação da árvore, e recursivamente processam os resultados das ramificações.

Essa estratégia de dividir para conquistar foi desenvolvida e refinada ao longo de vários anos por John Ross Quinlan. Sua contribuição inicial foi o algoritmo ID3, QUINLAN (1986). Várias melhorias foram realizadas nesse algoritmo, culminando no surgimento do algoritmo C4.5, QUINLAN (1993). Posteriormente surgiu o C5.0, uma versão comercial do C4.5.

Conforme CALINSKI et al. (1974) e STUART (2003), o uso de algoritmos de indução de árvores de decisão é bastante difundido em virtude das seguintes características:

- ✓ Gera modelos de simples interpretação. Uma decisão complexa (prever o valor da classe) é decomposto numa sucessão de decisões elementares.
- ✓ É um método não paramétrico.
- ✓ Faz a decomposição de problemas complexos em vários problemas simples.
- ✓ Não assume nenhuma distribuição particular para os dados.
- ✓ Pode construir modelos para qualquer função desde que o número de exemplos de treino seja suficiente.

Em contrapartida, algumas das dificuldades também são consideradas, como por exemplo:

- ✓ A análise da árvore é passo de mineração de dados (avaliação do modelo). O usuário deve avaliar como, quando e onde podar a árvore.
- ✓ Possibilidade de fragmentar conceitos.
- ✓ Em alguns casos pode retornar estruturas muito extensas.

QUINLAN (1986 e 1993) demonstra que os critérios de parada da divisão recursiva dos elementos do algoritmo de classificação por indução de regras de associação são os seguintes:

- ✓ Todos os indivíduos pertencem a mesma classe.
- ✓ Todos os indivíduos possuem os mesmos valores dos atributos, mas diferentes classes.
- ✓ O número de indivíduos é inferior a certo limite.

Utilizaremos nesta pesquisa o algoritmo C5.0 disponível no software SPSS Clementine para ajudar o usuário especialista na tomada de decisões sobre a avaliação dos padrões das transações de compras da empresa em questão, utilizando como base de dados um data mart criado especificamente para este fim. Maiores detalhes sobre este método, incluindo sua fundamentação matemática, pode ser visto em ZAKI et al. (1998) e STUART (2003).

## Algoritmo: C5.0 Node

Um método para elaboração de árvores de decisão e para geração de regras amplamente utilizado e incorporado por outras ferramentas para mineração de dados é o C5.0, que é implementado por um software com o mesmo nome, desenvolvido pelo Professor Ross Quinlan. Este método é uma evolução do método anterior, conhecido como ID3 (Iterative Dichotomizer 3). Por ter sua teoria descrita em livro (Quinlan, 1993) e ser disponibilizado com fontes, o método disseminou-se rapidamente e hoje é incorporado em várias ferramentas comerciais da categoria.

Os experimentos que faremos nesta pesquisa, utilizaram o algoritmo C5.0 presente no módulo de classificação do software Clementine, versão 11, e pode ser visto na Figura 29.

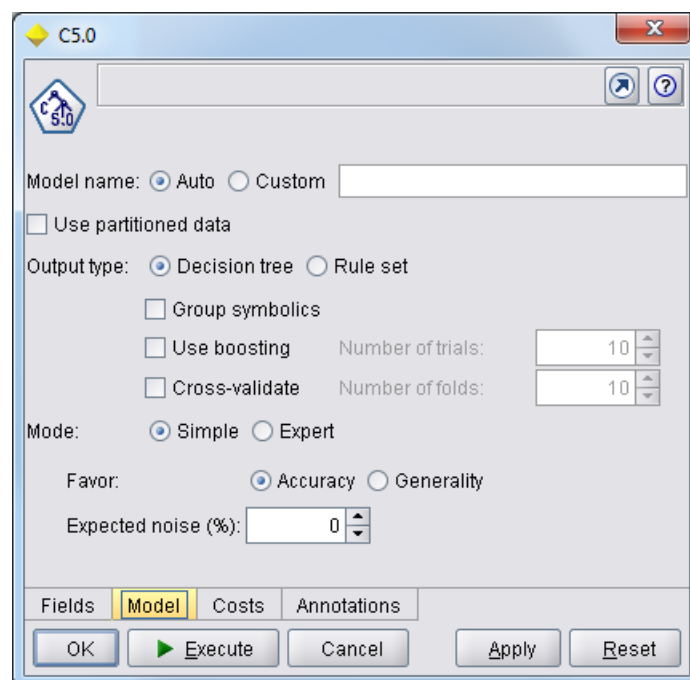


Figura 29 - C5.0 Clementine

Os parâmetros de ajustes do modelo que podem ser configurados pelo usuário no software Clementine serão detalhados a seguir:

- ✓ *Model name*: Especifica o nome do modelo que será produzido. Caso o usuário deixe selecionada a opção “auto”, o Clementine usará o nome da variável definida como consequência (target).
- ✓ *Output type*: Aqui deve se especificar que tipo de saída o algoritmo deve produzir: conjunto de regras (rule set) ou uma árvore de decisão (decision tree).
- ✓ *Group symbolics*: Com esta opção selecionada, o algoritmo agrupará as variáveis categóricas que tiverem padrões semelhantes no nível superior. Caso esta opção não seja selecionada, o C5.0 dividirá o nó pai (de nível superior) e criará um nó filho para cada variável categórica. Por exemplo: uma base de dados possui um campo de cor com três atributos (vermelho, verde e azul). Por padrão, o algoritmo cria um

grupo para cada atributo, mas se esta opção for selecionada e os registros onde cor=vermelho forem semelhantes aos registros de cor=azul, o C5.0 criará um único agrupamento para os registros verdes e agrupará todos os registros vermelho e azul.

- ✓ *Use boosting*: O algoritmo C5.0 tem um método especial para melhorar sua taxa de acerto, chamado de boosting (ou reforço). Ele funciona através da construção de vários modelos em seqüência. Ou seja, assim que o primeiro modelo é construído, outro é construído em seguida, de tal forma que ele incida sobre os registros que foram erroneamente classificados pelo primeiro modelo. Logo depois, um terceiro modelo é construído para se concentrar nos erros do segundo modelo, e assim por diante. No final, os registros são classificados através da aplicação do conjunto de modelos obtidos no processo, usando um procedimento de votação ponderada para combinar todas as previsões anteriores em uma previsão global. O reforço (boosting) pode melhorar significativamente a precisão de um modelo C5.0, mas também exige maior tempo de treinamento. A opção "*Number of trials*" permite que seja definido a quantidade de modelos que serão usados para criar o modelo reforçado.
- ✓ *Cross-validate*: A validação cruzada é uma importante ferramenta utilizada para a determinação de parâmetros em modelos estatísticos. Basicamente, a técnica consiste de dividir a série de dados utilizados em dois grupos: o período de treinamento ("grupo teste") e o período de validação. No primeiro, ajusta-se um modelo. No segundo, é avaliado o desempenho do modelo ajustado, segundo os dados de validação. Assim, compara-se ao final o desempenho de diferentes procedimentos específicos que modificam o modelo (número de preditandos, tipos de modelos, etc.).

Se essa opção for selecionada, o C5.0 usará os modelos construídos com os dados de treinamento para estimar toda a precisão do modelo construído nos dados do processamento. Isso é útil se o conjunto de dados for muito pequeno para dividir em conjuntos de teste. Os modelos de validação cruzada são descartados depois de calculado todas as estimativas de precisão. É possível especificar o número conjuntos de treinamentos, ou o número de modelos utilizados para validação cruzada.

- ✓ *Favor*: Por padrão o algoritmo tentará produzir a árvore com maior acuracidade possível. Em alguns casos pode ocorrer *overfitting* (o modelo é totalmente ajustado para os dados de treinamento e apresenta resultados ruins na base de dados de processamento). A opção "Generality" pode ser usada para deixar o modelo menos suscetível a esse problema.
- ✓ *Expected noise (%)*: Permite especificar a proporção esperada de dados com ruídos ou errados no conjunto de treinamento.

Após concluir a revisão bibliográfica, iniciaremos com a fase experimental, que seguirá a metodologia de Busca de Conhecimento em Banco de Dados (KDD – Knowledge Database Discovery).

### 3) Desenvolvimento da metodologia

Neste capítulo detalharemos a execução das seguintes etapas distintas do KDD: coleta dos dados, pré-processamento e processamento (ou mineração de dados).

#### 3.1) Coleta e consolidação dos dados

A empresa utilizada nesta pesquisa possui diversas bases de operações e fábricas espalhadas pelo Brasil. Sua matriz fica localizada no Rio de Janeiro. Todas as filiais utilizam o mesmo sistema ERP Sispro de forma independente; ou seja, existe uma base de dados em cada localidade. Ao final do mês, no período de fechamento contábil, as movimentações financeiras são replicadas para uma base central que consolida os livros fiscais da corporação.

Embora operem de forma independente, algumas informações precisam estar sincronizadas entre todas as filiais, como por exemplo, o cadastro de clientes e fornecedores. As tabelas de clientes e fornecedores do banco de dados de cada filial são replicadas diariamente, conforme o esquema de replicação que está representado na Figura 30.

A replicação permite que, caso uma das fábricas do Sul do país cadastre um novo cliente, em menos de 24 horas esta informação estará disponível para que as bases do Nordeste possam emitir uma nota fiscal para este novo CNPJ cadastrado. O mesmo acontece com fornecedores. Caso alguma filial qualifique negativamente um determinado fornecedor, nenhuma outra localidade poderá emitir pedidos para o CNPJ em questão.

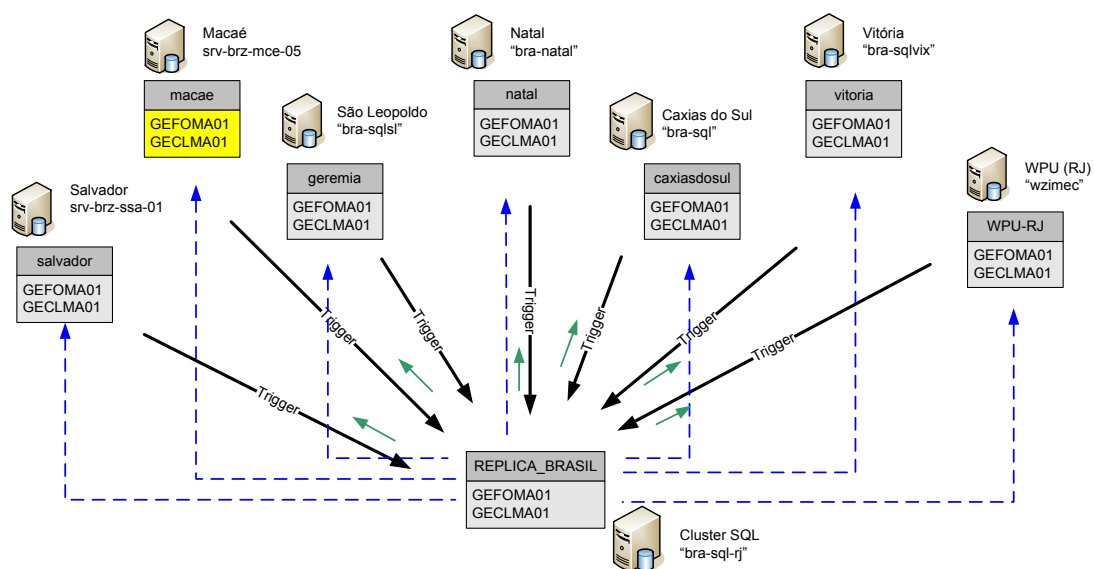


Figura 30 - Estrutura de replicação de banco de dados.

As informações de replicação são importantes, pois como trabalharemos os dados de transações de compras, a identificação correta dos fornecedores é fundamental.

Conforme dito anteriormente, utilizaremos os dados do módulo de Suprimentos do ERP Sispro, que permite o registro e acompanhamento dos processos de compras, recebimento, movimentação de estoques e emissão de notas de transferência, devolução e remessa. A base operacional de Macaé foi a escolhida para este experimento, e o período de dados disponibilizado para análise vai de 01 de janeiro de 2008 a 31 de dezembro de 2009.

Para que fosse possível realizar todos os experimentos sem interferir no ambiente de produção detalhado na Figura 30 e permitir que o usuário especialista manipulasse diretamente a massa de dados, foi criado um ambiente de desenvolvimento, também em Macaé, somente para fins de pesquisa. Este ambiente é isolado do ambiente de produção e foi configurado com uma cópia da base de dados com as informações do período de análise.

No ambiente de testes, temos então a cópia da base de dados do ERP chamada de REPLICA\_SISPRO. Esta base fica armazenada em um servidor diferente, mas utiliza a mesma versão do SQL Server (versão 2005) instalado em produção. A interação com o usuário especialista se deu, inicialmente, através do uso de uma aplicação "front-end" desenvolvida em Delphi especificamente para este estudo, que utiliza o componente OLAP conhecido como *PivotCube*.

O *PivotCube*, exemplificado na Figura 31, é um componente mantido pela empresa PivotWareLab e por ser baseado em OLAP possibilita análise multidimensional dinâmica dos dados através da navegação em níveis (conhecido nos sistemas de Business Intelligence como *Drill Down* e *Drill Up*). A grande vantagem identificada pelo uso deste componente foi poder extrair e integrar dados de múltiplas fontes, permitindo que a informação fosse analisada pelo usuário especialista de forma contextualizada e extremamente detalhada.

		Summa Value	Quantity	Summa 3rd Quatile	Summa	Quantity
2001		14,772,620.00	20601.00	598050.25	5,623.00	179.00
Spring		2,968,510.00	6405.00	523875.50	0.00	0.00
March		923,644.00	2332.00	245084.00	0.00	0.00
April		2,464,866.00	4073.00	632117.25	0.00	0.00
May		0.00	0.00	0.00	0.00	0.00
Summer		2,791,465.00	3885.00	537554.50	5,623.00	179.00
June		490,697.00	1003.00	225348.50	0.00	0.00
July		693,787.00	753.00	196211.00	5,623.00	179.00
August		1,646,981.00	2129.00	567237.75	0.00	0.00
Autumn		3,624,445.00	5233.00	568912.00	0.00	0.00
September		847,892.00	908.00	334406.00	0.00	0.00
October		1,517,370.00	2374.00	1020830.00	0.00	0.00
November		1,299,183.00	2051.00	495266.50	0.00	0.00

Figura 31 – PivotCube

A aplicação *PivotCube* permitiu que fosse criada diversas consultas à base de dados do experimento. Estas consultas, definidas em conjunto com o usuário especialista, foram feitas em linguagem SQL e armazenadas em duas “stored procedures” no banco de dados SQL Server.

O ambiente de configuração, após totalmente desenvolvido, pode ser visto na Figura 32 abaixo. Nesta figura temos o ambiente em produção e o desenvolvimento no mesmo esquema, ou seja, temos a indicação da base em produção do ERP Sispro e o REPLICA\_SISPRO, assim como a base de dados QUERY\_CUBE que é utilizada pela ferramenta *PivotCube*.

Pela interpretação da Figura 32, acredita-se ficar claro o entendimento da forma como foi feito a consolidação dos dados nesta etapa da pesquisa: ao iniciar a execução da ferramenta desenvolvida com o componente *PivotCube*, o usuário especialista define as variáveis que deseja utilizar no experimento – o objetivo é definir as mais importantes na caracterização das transações de compras da empresa. Os parâmetros definidos pelo especialista são enviados para as stored procedures do banco REPLICA\_SISPRO através de passagem de parâmetros. O comando SQL então é executado e, em seguida, o resultado é exibido na tela do usuário especialista conforme visto na Figura 31.

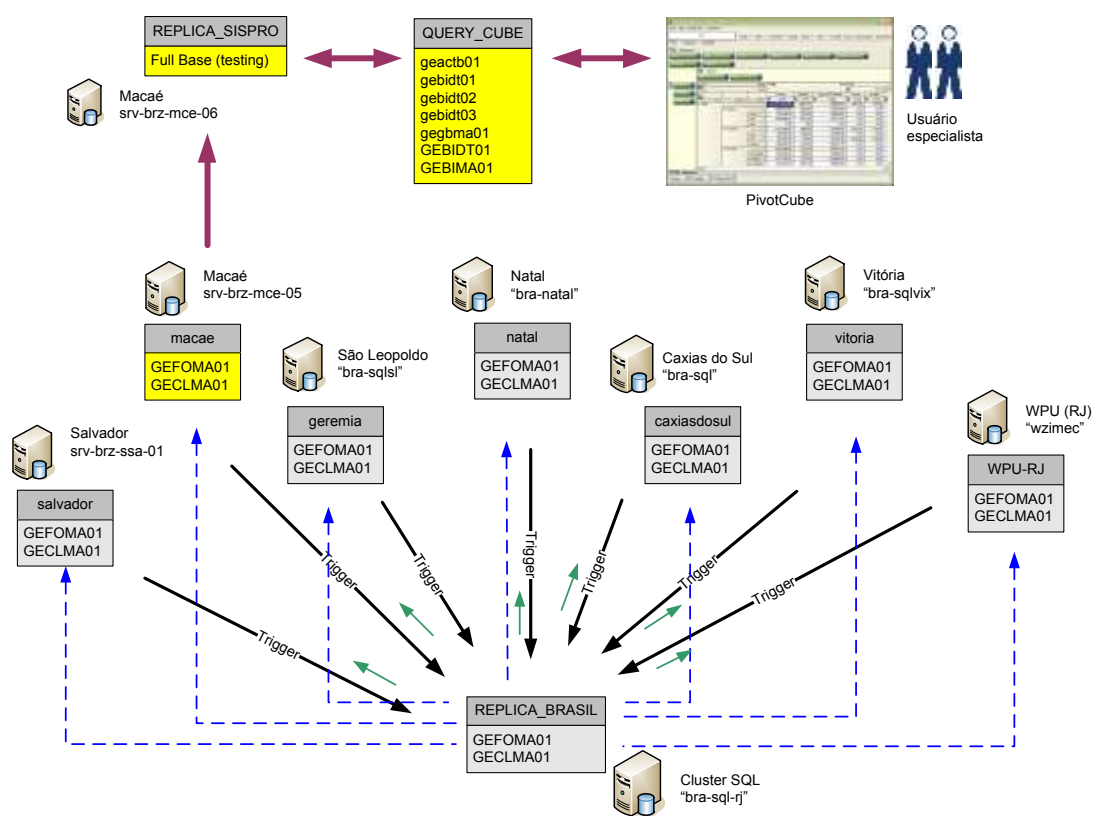


Figura 32 - Ambiente de desenvolvimento

As tabelas da base QUERY\_CUBE, utilizadas pela aplicação *PivotCube*, estão descritas na Figura 32 e o MER (Modelo Entidade Relacionamento) na Figura 33.

A tabela GEBIMA01 contém os parâmetros que podem ser configurados pelo especialista. Na Tabela 8 temos dois tipos de consultas configuradas, que fazem passagem de parâmetros para as *procedures* [relatorio\_Notas] e [RELATORIO\_

ENTRADAS]. Estas *procedures* estão disponíveis para consulta no Anexo I – Comandos SQL.

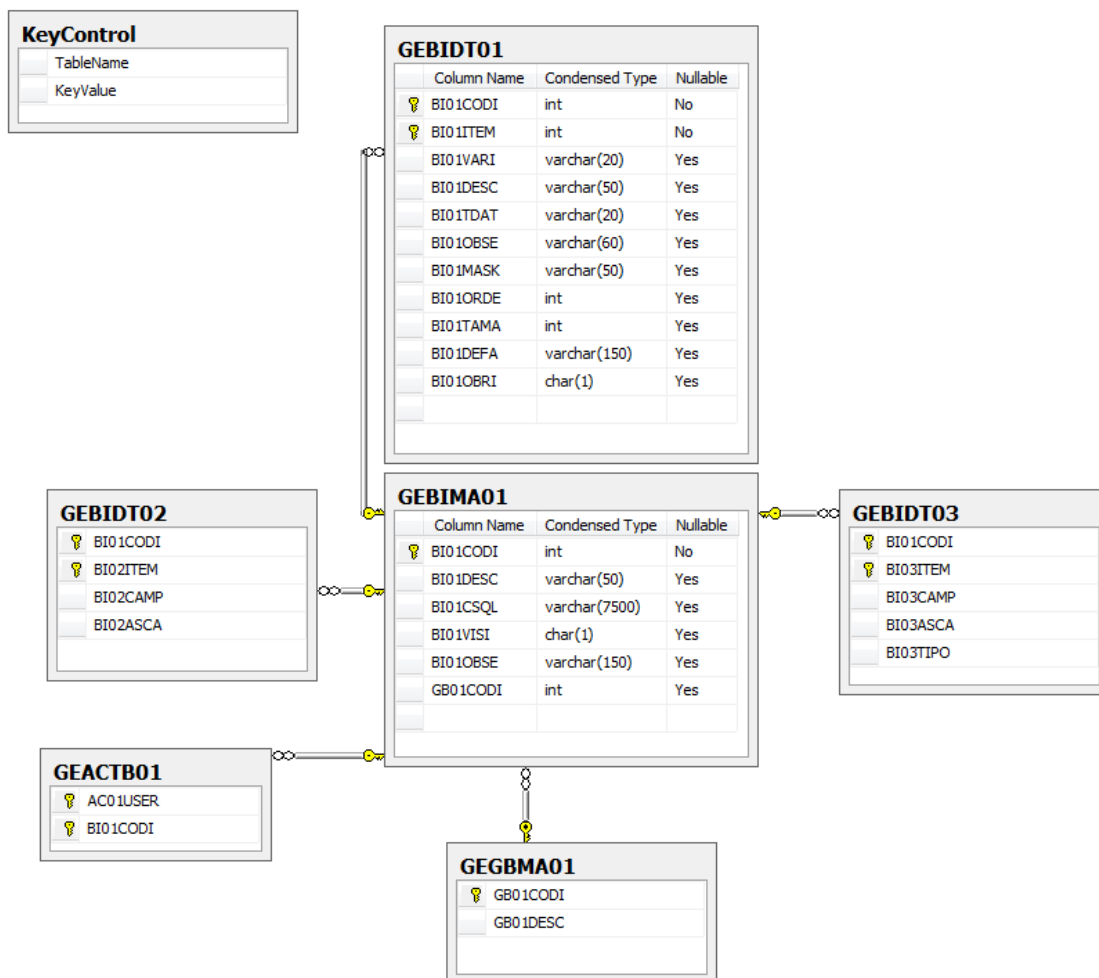


Figura 33 - Modelo Entidade Relacionamento da tabela QUERY\_CUBE

BI01CODI	BI01DESC	BI01CSQL	BI01VISI	BI01OBSE	GB01CODI
44	NOTAS SAIDA	exec [relatorio_Notas] S ""+@estab+"" , ""+@TEXTO +"" ,""+@NOTA +"" , ""+ @PV +"" , ""+@lw +"" , ""+ @CC +"" , ""+@dt_NFin+"" , ""+@dt_nFfim +""	S	NULL	5
48	NOTAS DE ENTRADA	exec [RELATORIO_ENTRADAS] S ""+@estab+"" , ""+@lw+"" , ""+@requisitante+"" , ""+@comprador+"" , ""+@codforn+"" , ""+@nomeforn+"" , , ""+@semdoc+"" , ""+@por+"" , ""+@documento+"" , ""+@dtini+"" , ""+@dtfim+""	S		5

Tabela 8 - PivotCube: passagem de parâmetros para as stored procedures

Todas as consultas configuradas no *PivotCube* foram feitas de forma a permitir a análise através do relacionamento de todas as variáveis utilizadas no módulo de Suprimentos do ERP Sispro – Tabela 9.

Estabelecimento	Requisição	Situação	Requisitante	Nome Requisitante
Linha de Negócio	Cc	Fdc	Contrato	Departamento
Goa/afe	Security Number	Data Rc	Data Necessidade	Aprovador
Data Aprovação	Nome Aprovador	Item	Quantidade	Código
Descrição Material	Sutuação Atendimento	Cotação	Data Inclusão Cotação	Data Final da Cotação
Situação Cotação	Pedido	Data Pedido	Previsão de Entrega	Comprador
Nome Comprador	Item Pedido	Unidade Medida	Quantidade Pc	Valor Unitário Pc
Vl_total	Fornecedor Pc	Razão Fornecedor Pc	Vencido (S/N)	Folow Up 1
Folow Up 2	Folow Up 3	Tipo Nota	Nota Fiscal	Fornecedor
Razão Fornecedor	Tipo Recebimento	Cnpj	Uf	Pais
Data Nota	Data Entrada	Data Cadastro Nota	Data Alteração Nota	Usuário
Nome Usuário	Item	Produto	Descrição	Ncm
Tipo Material	Depósito	Unidade de Medida	Quantidade	Valor Unitário
VL_TOTAL_ITE M	Conta Contábil	Custo	Base Icms	Icms
Valor Contábil	Diferencial Icms	Base Ipi	Ipi	Pis
Cofins	Iss	Inss	Rateio Despesas	Tipo Frete
Operação Fiscal	Cfop	Descrição Operação Fiscal	Vencimento	Valor Fatura
faturaprop	Grupo Contábil	Cst	Notas Relacionadas	Última Nota de Saida
Item da Ultima Nota	Data da Ultima Nota	Documento	Ano	Mês

Tabela 9 - Variáveis envolvidas nas transações de compras

Terminado os ajustes com as consultas ao banco de dados, é necessário realizar o tratamento dos dados, para que posteriormente, na etapa de processamento, sejam aplicados os algoritmos de mineração de dados.

## Tratamento da base de dados

De acordo com BISHOP (2006) e MARKOV et al. (2007), este é o processo mais demorado no KDD (Knowledge Discovery in Database), pois é feito o tratamento de valores ausentes, verificado inconsistência e redundância de dados, e visto outros problemas mais específicos que podem alterar o resultado do algoritmo de mineração.

Após a consolidação dos dados, foram obtidos 27.618 registros, que compõem todas as requisições feitas internamente ao setor de compras e os pedidos emitidos para os fornecedores. Detalharemos a seguir todo o tratamento feito na base de dados em conjunto com o especialista.

Inicialmente, por utilizarmos o histórico de uma base de dados em produção, foi preciso filtrar o estado das transações (campo “**Situação**”) de acordo com o interesse do estudo. Segundo o usuário especialista, transações canceladas, suspensas ou totalmente pendentes (entenda totalmente pendente como sendo uma ordem de compra com diversos itens sendo que nenhum deles foi atendido até o momento da extração dos dados) não serão consideradas nesta pesquisa.

- Esta seleção eliminou 50 transações da base de dados. Observação: foram mantidas todas as transações pendentes onde a ordem de compra possui um ou mais itens e que tenha sido parcialmente atendida.

Outra questão importante tem haver com a estrutura organizacional da corporação. A empresa possui diversas Linhas de Negócio (campo “**LinhadeNegócio**”). Cada Linha de Negócio possui um centro de custo (campo “**Cc**”) de onde se debitam as despesas. Todas as transações devem pertencer a uma Linha de Negócio. Casos que a Linha de Negócio ou o centro de custo não foi informado foram eliminados da base de dados.

- Esta seleção eliminou 103 transações da base de dados.

No banco de dados, todas as requisições de compras (RC) não aprovadas possuem o campo “**Aprovador**” nulo. As RCs com valor nulo para esta variável foram desconsideradas.

- Esta seleção eliminou 9.116 requisições da base de dados.

Em relação ao cadastro de materiais, todo item possui um código único cadastrado no sistema ERP. As solicitações e transações de compra de materiais que foram feitas utilizando o **código** número 999999999 (material genérico) e 999999998 (serviço genérico) foram desconsideradas.

- Esta seleção eliminou 717 transações da base de dados.

Ainda em relação ao código de materiais, identificamos requisições de compras em aberto, ou seja, o requisitante não concluiu a digitação informando o código do produto, mas deixou o processo salvo no sistema, o que acabou gerando um número de identificação para o processo. Todas as transações que o **código** do material não foi informado foram desconsideradas.

→ Esta seleção eliminou 114 requisições da base de dados.

Por orientação do usuário especialista não foi considerada nesta pesquisa a contratação de serviços, somente compra de ferramentas ou materiais de consumo. Com isso, todos os campos "**Tipo Material**" igual a serviço ou nulo foram desconsiderados.

→ Esta seleção eliminou 5.788 transações da base de dados.

As variáveis "**Tipo Material**" e "**Depósito**" estão relacionadas. Todos os bens adquiridos que possuem o campo "**Depósito**" nulo são considerados pelo ERP Sispro como bem imobilizado<sup>14</sup>. Como o objetivo deste estudo é identificar padrões de compras de materiais de uso e consumo, todas as transações de compra que tiverem materiais com o campo "**Depósito**" nulo serão desconsideradas.

→ Esta seleção eliminou 1.512 transações da base de dados.

Após a etapa de tratamento dos dados, foram eliminados 17.400 registros do banco de dados REPLICA\_SISPRO. Os 10.218 registros que restaram, serão utilizados nas próximas etapas do processo KDD.

---

<sup>14</sup> Bem imobilizado é um conceito empregado na contabilidade para definir quais são os bens que não sofrem movimentação constante. São essenciais para a empresa continuar operando e não podem ser convertidos em dinheiro imediatamente. Alguns exemplos são: os imóveis, os equipamentos, os utensílios, as ferramentas, as patentes e etc.

### 3.2) Seleção e pré-processamento

Este capítulo aborda as atividades realizadas na preparação dos dados antes que sejam aplicados os algoritmos de mineração de dados.

Na seleção de atributos, conforme visto em HAN et al. (2001) e o próprio nome já diz, o objetivo é escolher um subconjunto de atributos (ou variáveis) ou até mesmo criar outros atributos que substituam um conjunto de variáveis a fim de reduzir a dimensão do banco de dados. Com essa redução de dimensão, reduz-se a complexidade do banco de dados e assim o tempo de processamento para extrair dele algum conhecimento. Além disso, atributos desnecessários podem causar ruído no resultado final e isto pode ser evitado com a aplicação de técnicas de Seleção de Atributos.

Segundo KOHAVI et al. (1997), existem duas abordagens principais para se realizar a seleção de atributos: *filtro* e *wrapper*. KOHAVI et al. (1997) coloca que nos *filtros* os atributos são ignorados independentemente dos valores contidos nos registros do conjunto de teste, ou seja, apenas as propriedades dos atributos são consideradas para classificá-la como útil ou não; neste caso os atributos são todos selecionados independente da execução do algoritmo de mineração de dados. A seleção dos atributos é geralmente feita pelo especialista e ocorre de maneira estática.

Na segunda abordagem (*wrapper*), a seleção dos atributos também independe do algoritmo de mineração, porém é feita a cada iteração. Novos subconjuntos de atributos são escolhidos e avaliados dinamicamente enquanto a mineração ocorre até que não seja viável realizar mais melhorias na escolha dos atributos, segundo algum critério de avaliação. Neste caso, as variáveis ou conjuntos de variáveis são avaliados considerando os valores encontrados nas saídas dos algoritmos de mineração de dados aplicados no conjunto de teste.

A etapa de seleção de atributos como pode ser vista, está presente apenas no início do processo de mineração de dados, porém sua aplicação requer grande responsabilidade, pois toma a maior parte do processo de mineração e consome muito processamento.

O pré-processamento engloba diversas técnicas de organização, tratamento e preparação dos dados. É uma etapa que possui fundamental relevância no processo de descoberta de conhecimento – KDD (Knowledge Discovery in databases). Segundo HAN et al. (2001), compreende desde a correção de dados errados até o ajuste da formatação dos dados para os algoritmos de mineração de dados que serão utilizados. Abaixo temos algumas das atividades feitas na base de dados REPLICA\_SISPRO que fazem parte do processo de pré-processamento:

- ✓ Binarização – alguns algoritmos utilizados na extração de regras de associação requerem que os dados estejam na forma de atributos binários. Assim, muitas vezes tanto os atributos contínuos quanto os discretos necessitam ser transformados em um ou mais atributos binários.
- ✓ Construção de atributos – essa operação consiste em gerar novos atributos a partir dos atributos existentes. A importância desse tipo de operação é justificada, pois novos atributos, além de expressarem relacionamentos conhecidos entre atributos

existentes, podem reduzir o conjunto de dados simplificando o processamento dos algoritmos de mineração de dados.

- ✓ Discretização - alguns algoritmos, especialmente os algoritmos de classificação, requerem que os dados estejam na forma de atributos categorizados. Assim, muitas vezes é necessário transformar um atributo contínuo em categórico.

Ainda segundo HAN et al. (2001), a estratégia de simplificação dos dados deve atentar também para a questão da supressão de um atributo, que no caso da mineração de dados é muito mais delicado que a supressão de uma linha. Retirar atributos relevantes ou permanecer com atributos irrelevantes pode implicar na descoberta de padrões de baixa qualidade.

A seleção de atributos para este experimento se deu através do uso de ambas as abordagens: *filtro* e *wrapper*. Originalmente o usuário especialista definiu um conjunto de variáveis que, após processadas pelos algoritmos de mineração, não retornaram resultados expressivos. Daí então, as informações obtidas com os resultados anteriores motivaram a binarização de alguns atributos e, conseqüentemente na construção de novas variáveis. No final do processo, foram criados dois bancos de dados específicos com informações relativas aos materiais adquiridos pela empresa durante o período analisado e dados de caracterização das transações de compras. Estes bancos de dados com dados históricos específicos são popularmente conhecidos por *data marts*.

Algumas considerações importantes foram feitas pelo especialista durante a fase de seleção de atributos, principalmente após visualizar graficamente a distribuição das variáveis “Fornecedor Pc”, “Cfop”, “Uf” e “Comprador”, disponível na Figura 34.

- (a) As transações de compras de 2008 e 2009 utilizaram 539 fornecedores distintos, sendo 80,87% das transações feitas por 102 fornecedores. A grande quantidade de fornecedores impede negociações mais agressivas com os fornecedores e não permite alinhamento estratégico entre empresas.
- (b) Todas as transações foram caracterizadas por códigos CFOP (Códigos Fiscais de Operações e Prestações) que indicam a compra de material para uso ou consumo. Estes códigos representam os motivos reais da entrada ou saída de materiais e é usado pelo Governo para a cobrança de impostos.
- (c) Mais de 92% das aquisições foram feitas com fornecedores do eixo Rio – São Paulo, sendo que o grande volume está no estado do Rio de Janeiro. Visibilidade também para os processos de importação – fornecedores no Exterior representam 4,48% de todas as transações realizadas no período. Dado importante para o setor logístico da empresa.
- (d) Conforme visto no capítulo 2.1 (Evolução do setor de Compras) a empresa que vive na primeira fase de evolução possui funcionários com desempenho relacionado apenas à quantidade de pedidos emitidos, sem nenhuma preocupação com o controle de custos e redução de despesas. É o que podemos ver ao analisar que 50% de todas as transações foram feitas por apenas dois Compradores.

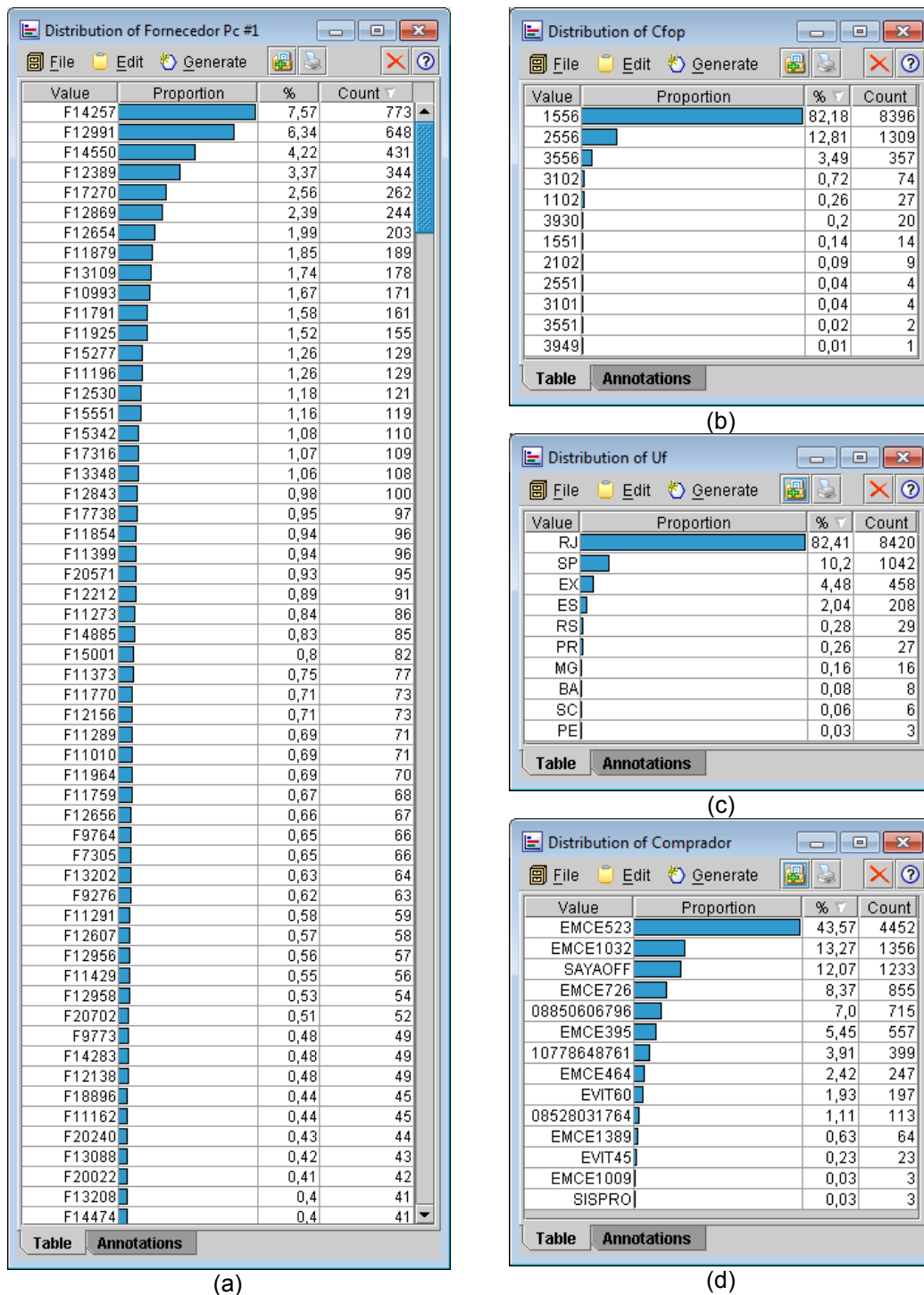


Figura 34 - Variáveis “Fornecedor Pc”, “Cfop”, “Uf” e “Comprador”

Analisando a Tabela 9 - Variáveis envolvidas nas transações de compras, podemos identificar as 95 (noventa e cinco) variáveis selecionadas inicialmente pelo especialista para caracterizar as transações de compras. Após exaustivo trabalho de análise de correlação de dados, foi possível diminuir a dimensionalidade do problema para aproximadamente 20% do total, incluindo os atributos que foram criados a partir dos atributos existentes. As variáveis utilizadas compõem os dois *data marts* criados, que são detalhados logo em seguida.

## Data Mart Pedido de Compras

Conforme visto em MONTERIO et al. (2004), a necessidade de analisar grandes volumes de dados levou os desenvolvedores a criarem ambientes que suportassem, de forma otimizada, os conceitos de multidimensionalidade de dados e navegabilidade hierárquica facilitada. Este ambiente introduz o conceito de *Business Intelligence*, que é utilizado nesta pesquisa através do uso do componente em Delphi *PivotCube*. Esta ferramenta oferece meios de organizar processos de extração de conhecimento explorando, principalmente, o potencial da tecnologia de Data Warehouse, representada na Figura 35 pela base de dados REPLICA\_SISPRO.

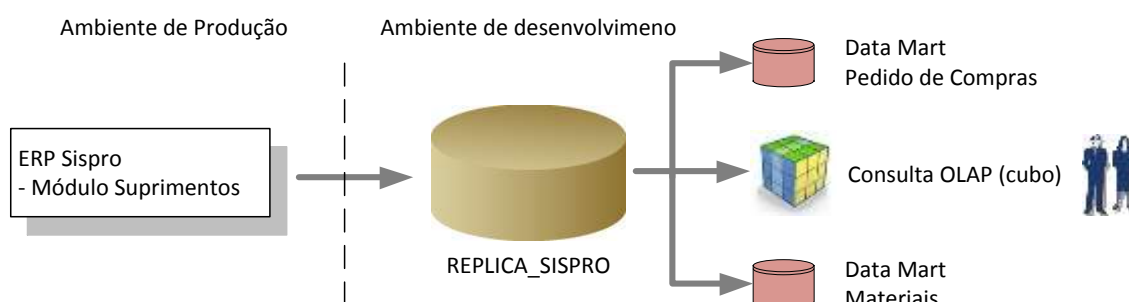


Figura 35 - Estrutura de dados

Os processos de extração de conhecimento desta pesquisa não foram realizados com os dados brutos da base REPLICA\_SISPRO, mas com dois conjuntos flexíveis de dados apresentados em um modelo dimensional que representa os dados de acordo com o processo de negócio: os data marts Pedido de Compras e Materiais.

MONTERIO et al. (2004) coloca que os data marts podem ser construídos através de bancos de dados relacionais ou bancos de dados OLAP – Online Analytical Processing (Cubos), mas que as consultas OLAP são muito mais rápidas e permitem maior interatividade dos usuários, se comparado às consultas Relacionais. Detalharemos a seguir os atributos da base (data mart) Pedido de Compras, que estão relacionados abaixo na Tabela 10.

1	<b>Requisição (primary key)</b>	8	contarcnpj	15	DataAprovação
2	Aprovador	9	contardeltem	16	DataEntrada
3	Cc	10	ContardeNcm	17	DataNota
4	Comprador	11	prazoaprovaçãocompra	18	DataPedido
5	Requisitante	12	prazochegadaterial	19	DataRc
6	TipoFrete	13	prazoemissãopedido	20	SomadeVL_TOTAL_ITEM
7	UnidNegocios	14	prazoliberaçãomaterial		

Tabela 10 - Atributos do Data Mart Pedido de Compras

O campo "**Requisição**" é a chave primária da tabela. A chave primária, também chamada de primary key ou simplesmente PK pelos SGBDs (sistemas gerenciadores de banco de dados), tem a função de tornar um registro em uma tabela único, de modo com que ele nunca se repita. Esta chave pode ser usada como um índice para os demais

campos da tabela e não pode conter valores nulos. Neste atributo temos 2.979 registros, que embora sejam numéricos, serão tratados como categóricos. Na Figura 36 temos a quantidade de requisições feitas em função do tempo.

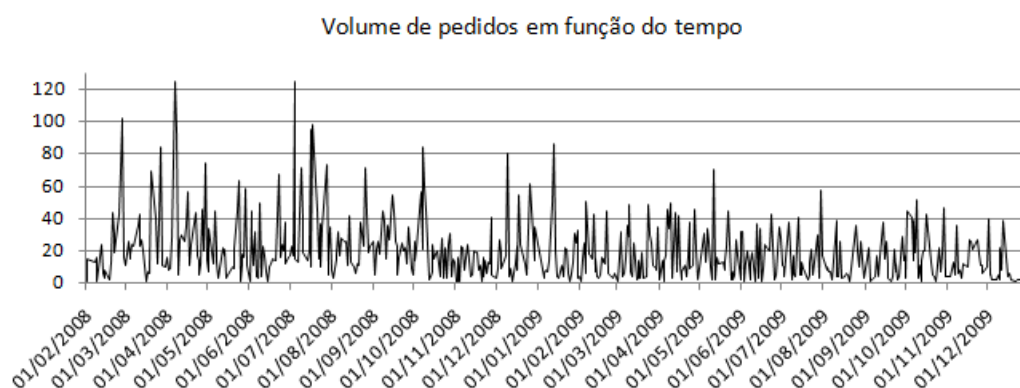


Figura 36 - volume de pedidos em função do tempo

Conforme visto anteriormente no capítulo 2.2, no processo de compra uma solicitação (requisição) mesmo contendo diversos itens só pode ser feita por um único requisitante. Da mesma forma, esta requisição é avaliada por um único aprovador e estará associada a apenas uma unidade de negócio e seu respectivo centro de custo. Caso aprovada, todos os itens da requisição serão submetidos para o mesmo comprador. Na Tabela 11 temos um resumo da quantidade de valores distintos que cada atributo possui na base de dados. Vale ressaltar que todas estes atributos são categóricos.

Atributo	Quantidade de valores distintos
Aprovador	54
Cc	51
Comprador	14
Requisitante	75
UnidNegocios	13

Tabela 11 - Variáveis e sua dimensionalidade

O atributo "**TipoFrete**" possui apenas dois valores categóricos: C (para CIF) ou F (para FOB). Estas designações são oriundas do inglês e significam:

- (i) CIF - Cost, Insurance and Freight: custo, seguro e frete e
- (ii) FOB – Free on Board: Livre a bordo.

Conforme o Art. 10 parágrafo 4º incisos I e II do RCTE (Regulamento do Código Tributário do Estado):

*I - FOB é aquele em que o custo do frete e do seguro é de responsabilidade do comprador ou destinatário;*

*II - CIF é aquele em que o custo do frete e do seguro é de responsabilidade do vendedor ou remetente.*

Na Tabela 12 temos a distribuição da variável "**TipoFrete**" durante o período analisado.

Tipo de Frete	Total	Total (%)
C (CIF)	684	23%
F (FOB)	2.295	77%
Total geral	2.979	

Tabela 12 - Tipos de frete

Com o objetivo de melhor caracterizar as requisições de compras, o usuário especialista sugeriu inicialmente a criação de três atributos numéricos: "**contarcnpj**", "**contardeItem**" e "**contardeNcm**".

O primeiro indica quantos fornecedores diferentes atenderam a requisição de compra, e o segundo, a quantidade de itens distintos que foram solicitados. Neste caso, a quantidade não é considerada, pois é comum encontrar pedidos com itens que utilizam diferentes unidades de medida (como metro, quilo, centímetro, litro, etc.). O objetivo do terceiro atributo é comparar seus valores com a quantidade de itens, variável "**contardeItem**".

Somente para informação, NCM é a sigla de Nomenclatura Comum do MERCOSUL, adotada pelos países: Brasil, Argentina, Paraguai e Uruguai desde janeiro de 1995, e tem por base o Sistema Harmonizado. O Sistema Harmonizado foi criado para facilitar as relações internacionais e estabelece uma padronização para classificação de mercadorias em 177 países. As informações de todas as variáveis numéricas podem ser vistas na Tabela 14.

Para que fosse possível calcular o tempo de atendimento das etapas relacionadas ao processo de compras, as seguintes variáveis foram utilizadas (todas do tipo data, no formato DD/MM/AAAA):

- ✓ **DataRc** – data que o solicitante emitiu a requisição de compra e submeteu para aprovação.
- ✓ **DataAprovação** – data que a requisição de compra foi aprovada.
- ✓ **DataPedido** – data que o comprador finalizou o processo de negociação e enviou a ordem de compra para o fornecedor.
- ✓ **DataNota** – data que o fornecedor emitiu a nota fiscal do produto solicitado pelo comprador.
- ✓ **DataEntrada** – data em que o produto chegou na empresa e o material disponibilizado no almoxarifado.

Com as informações dos atributos acima, foi possível criar outros atributos numéricos, medidos em dias, que expressassem os seguintes prazos:

- ✓ Prazo de aprovação da requisição de compra (atributo "**prazoaprovaçãocompra**"): indica o tempo que o aprovador levou para analisar a real necessidade da compra e autorizar o pedido. Este prazo é calculado da seguinte forma:

$$\text{prazoaprovaçãocompra} = \text{DataAprovação} - \text{DataRc}$$

- ✓ Prazo de emissão do pedido (atributo "**prazoemissãodopedido**"): indica o tempo que o comprador levou para solicitar e receber todas as cotações, realizar as negociações com os fornecedores e emitir o pedido. Este prazo é calculado da seguinte forma:

$$\text{prazoemissãodopedido} = \text{DataPedido} - \text{DataAprovação}$$

- ✓ Prazo de recebimento do material (atributo "**prazochegadamatériau**"): indica o tempo que o fornecedor levou para atender o pedido de compra e entregar a mercadoria. Este prazo é calculado da seguinte forma:

$$\text{prazochegadamatériau} = \text{DataNota} - \text{DataPedido}$$

- ✓ Prazo de liberação do material (atributo "**prazoliberaçãodomatériau**"): indica o tempo que a mercadoria leva para ser inspecionada pelo setor de qualidade e entregue ao setor de almoxarifado para que o requisitante possa retirar o produto. Este prazo é calculado da seguinte forma:

$$\text{prazochegadamatériau} = \text{DataNota} - \text{DataPedido}$$

O atributo "**SomadeVL\_TOTAL\_ITEM**" também é numérico e representa o valor total do pedido de compra. É a soma dos valores unitários de cada item multiplicados pela quantidade solicitada. Conforme dito anteriormente, as informações de todas as variáveis numéricas podem ser vistas na Tabela 14.

Após detalhar as variáveis juntamente com o usuário especialista, algumas observações pertinentes ao perfil de compra da empresa foram registradas. Abaixo as que mais se destacaram para este banco de dados nesta etapa do estudo:

- ✓ Há necessidade de melhorar a avaliação da necessidade de compra de um item e, também, de diminuir o tempo de aprovação das requisições de compra. Foi encontrado requisições com 255 dias de espera para aprovação do pedido. Isto gera retrabalho para a equipe de Compras, pois todo o processo de cotação deverá ser refeito.
- ✓ Verificamos que as requisições de compras possuem em média três itens. O máximo de itens distintos de uma solicitação são 78 produtos, relacionados a 23 códigos NCM, ou seja, diversas especificações do mesmo tipo de produto.

## Data Mart Materiais

Esta base de dados possui doze atributos, listados na Tabela 13, sendo a maioria calculada a partir de informações extraídas da base REPLICA\_SISPRO. Segundo o usuário especialista, o objetivo é caracterizar o consumo dos materiais cadastrados no sistema através das variáveis disponíveis no módulo de suprimentos do ERP Sispro.

<b>codmaterial (primary key)</b>	totaldeaprovadores
qtdecomprada	totaldecc
qtdefornecedores	totaldecompradores
qtdencm	totaldeunidnegocios
qtdeuf	totalrequisições
vl_total_pedido	totalrequisitantes

Tabela 13 - Atributos do Data Mart Materiais

Assim como no anterior, o primeiro atributo, "**codmaterial**", é a chave primária, ou primary key. Este valor é único e não possui valor nulo. A base conta com 5.738 itens cadastrados.

Os próximos atributos são numéricos e fornecem idéia de quantidade. São eles:

**qtdecomprada** – esta variável soma todas as quantidades compradas para um determinado código de material durante o período analisado. O valor deste campo independe da unidade de medida, seja centímetro, metro, quilo, litro e etc., pois o valor do atributo é simplesmente numérico.

**qtdefornecedores** – indica o número de fornecedores diferentes que forneceram um determinado material ao menos uma vez durante o período analisado. As informações das variáveis numéricas podem ser vistas na Tabela 14.

**qtdencm** – este campo foi utilizado pelo usuário especialista para se certificar que cadastro de materiais está correto, pois cada código de material deve estar relacionado a apenas um código NCM.

**qtdeUF** – indica a quantidade de Unidades Federativas diferentes que forneceram um determinado material ao menos uma vez durante o período analisado. Este campo também é utilizado quando o material é importado.

**totaldeaprovadores** – agrega num único campo a quantidade de pessoas que aprovou requisições de compras com o referido material na lista de itens solicitados.

**totaldecc** e **totaldeunidnegocios** – indica quantos centros de custo e unidades de negócio já utilizaram um determinado material.

**totaldecompradores** – indica quantos compradores já negociaram um determinado tipo de material.

**totalrequisições** – informa a quantidade de requisições feitas para um material em específico.

**totalrequisitantes** – grega num único campo a quantidade de pessoas emitiram requisições de compras com o referido material na lista de itens solicitados.

O atributo “**vl\_total\_pedido**”, diferentemente do data mart Requisições Compras, possui o somatório de valores gastos na compra de um único material específico. No caso anterior, todo o pedido era considerado, agora, somente o valor e quantidade do material.

Após detalhar todas as variáveis juntamente com o usuário especialista, algumas observações pertinentes ao perfil de compra da empresa foram registradas. Abaixo as que mais se destacaram nesta etapa do estudo:

- ✓ Até o final de 2009 tínhamos cadastrado 5.738 itens. Destes, 4.163 itens (72,6%) foram adquiridos através de compra única, e 1.575 itens (27,4%) possuíam mais de uma requisição de compra cadastrado no sistema ERP Sispro. Por isso, a grande maioria dos pedidos de compras foram emitidos para apenas 1 fornecedor (CNPJ), mas há casos em que até 6 fornecedores distintos foram utilizados para atender o pedido.
- ✓ A variação nula na quantidade de códigos NCM por material indica que não temos erros no cadastro de materiais. Cada código de material está atribuído a apenas um NCM.
- ✓ Houve casos do mesmo código de material ter sua compra autorizada por onze aprovadores e ser negociado por sete compradores distintos que nem sempre utilizaram o mesmo fornecedor. Isto é um ponto extremamente negativo, pois é necessário ter volume de compra com os fornecedores para negociar preços mais agressivos.
- ✓ Os materiais que são utilizados por mais de uma unidade de negócio devem ser mantidos no estoque central, a fim de que sejam centralizados todas as requisições de compras num único solicitante, aprovador, comprador e fornecedor. Assim, é possível manter boa relação com o fornecedor, além de possibilitar menores prazos de entrega e menor custo por item.
- ✓ O departamento de Compras deve ficar atento ao caso de materiais que são adquiridos em mais de uma UF, visto que o custo do frete influencia no custo médio do produto.

Data Mart REQ COMPRAS	Count	Mean	Min	Max	Range	Variance	Standard Deviation	Standard Error of Mean
contar cnpj	2.979	1,13	1	6	5	0,25	0,50	0,01
contar de Item	2.979	3,43	1	78	77	25,75	5,08	0,09
Contar de Ncm	2.979	1,77	1	23	22	2,94	1,71	0,03
prazo aprovação compra (dias)	2.979	3,79	-	255	255	148,93	12,20	0,22
prazo chegada material (dias)	2.979	6,17	- 3.282	249	3.531	4.038,85	63,55	1,16
prazo emissão do pedido (dias)	2.979	13,84	- 235	225	460	441,37	21,01	0,39
prazo liberação do material (dias)	2.979	5,94	-	3.287	3.287	3.733,75	61,10	1,12
Soma de VL_TOTAL_ITEM	2.979	R\$ 5.889,40	R\$ 6,00	R\$ 1.022.129,34	R\$ 1.022.123,34	R\$ 1.127.957.925,28	R\$ 33.585,09	R\$ 615,33

Data Mart Materiais	Count	Mean	Min	Max	Range	Variance	Standard Deviation	Standard Error of Mean
totalrequisições	5.738	1,78	1	41	40	5,78	2,40	0,03
totalrequisitantes	5.738	1,22	1	11	10	0,39	0,63	0,01
totaldeaprovadores	5.738	1,31	1	11	10	0,60	0,78	0,01
totaldecompradores	5.738	1,22	1	7	6	0,33	0,57	0,01
totaldeUnidNegócios	5.738	1,09	1	6	5	0,13	0,36	0,01
totaldeCC	5.738	1,15	1	11	10	0,28	0,53	0,01
qtdeNCM	5.738	1,00	1	1	-	-	-	-
qtdeFornecedores	5.738	1,23	1	12	11	0,43	0,65	0,01
qtdeUF	5.738	1,04	1	4	3	0,04	0,20	0,00

Tabela 14 - Data Marts: REQ COMPRAS e Materiais

### 3.3) Processamento – Mineração de dados

Segundo STUART et al. (2003), a mineração de dados é a etapa do processo de KDD que consiste em aplicar técnicas e algoritmos de mineração onde, dentro de tempos aceitáveis, produzam um enumerado de padrões (ou modelos) úteis sobre os dados analisado. Ao final dessa etapa a ferramenta pode gerar um conjunto de descobertas, que depois de interpretadas pelo usuário especialista se transformam em conhecimento.

Existe uma grande variedade de técnicas de data mining, classificadas em diversos critérios, cada uma delas com características próprias que definem sua aplicabilidade. MARKOV (2007) coloca que certamente não existe uma única técnica que resolva todos os problemas. A familiaridade com as técnicas existentes é importante para proporcionar a melhor abordagem para o problema apresentado e as características do contexto de aplicação.

Portanto, conforme WITTEN et al. (2005) para cada tipo de aplicação deve-se utilizar um conjunto de algoritmos com o objetivo de extrair padrões e relações dentro de uma base de dados que se adéque ao caso. As ferramentas de data mining existentes incluem o uso de algoritmos de classificação, clusterização (análise de agrupamentos), extração de regras de associação, entre outros.

Neste experimento, por exemplo, em função das características dos dados, optou-se por aplicar os métodos de clusterização e extração de regras de associação. Os algoritmos utilizados foram: K-means, Two-step Node, C5.0 Node, GRI Node e Apriori Node, do software Clementine (versão 11.1 para o sistema operacional Windows).

Detalharemos primeiramente os resultados obtidos com os algoritmos de agrupamento Two-step cluster e k-means, utilizados no data mart Materiais (DM Materiais). Os algoritmos de mineração de dados foram executados diversas vezes e, a cada iteração com o usuário especialista, atributos diferentes foram utilizados.

Data Mart Materiais	
Variáveis da base de dados (data mart)	Variáveis utilizadas para clusterização
codmaterial	totalrequisições
totalrequisições	totalrequisitantes
totalrequisitantes	totaldeaprovadores
totaldeaprovadores	totaldecompradores
totaldecompradores	totaldeUnidNegocios
totaldeUnidNegocios	qtdeFornecedores
totaldeCC	qtdeUF
qtdeNCM	
qtdeComprada	
qtdeFornecedores	
qtdeUF	
VL_TOTAL_PEDIDO	
Total 12	Total 7

Tabela 15 - DM Materiais – variáveis utilizadas na clusterização

O conjunto de variáveis que apresentou melhor resultado de análise está descrito na Tabela 15. Abaixo temos algumas observações feitas sobre os atributos do banco de dados DM Materiais:

- ✓ A variável *totaldeCC* foi desconsiderada em função do interesse de realizar uma abordagem mais ampla, por unidades de negócio e não por centros de custo.
- ✓ O campo *qtdeComprada* possui diferentes unidades de medida (unidade, metro, quilo, litro, etc.), o que impossibilitaria a análise correta das medidas de distância de cada indivíduo da base.
- ✓ As variáveis *vl\_total\_pedido* e *qtdencm* tiveram maior utilidade quando usadas na extração de regras de associação.
- ✓ Por último, temos a variável *codmaterial*, que também foi descartada por ser a única categórica do conjunto utilizado para clusterização.

### Algoritmo: Two-step cluster

O método de agrupamento de dois estágios permite agrupar dados sem que seja necessário informar ao algoritmo o número de classes. Esta é um grande diferencial quando não se tem a idéia da quantidade de grupos possíveis.

Segundo o manual do usuário do Clementine, um pré-requisito para a execução do algoritmo é preencher o valor do campo "specify number of clusters". Neste caso, o valor mínimo e máximo foram 2 e 15, respectivamente.

Foram feitas diversas execuções com os dados do data mart Materiais, comparando as saídas dos dados normalizados e comparando com os dados sem normalização.

Os resultados da clusterização podem ser vistos na Figura 37, que mostra apenas os valores centrais de cada variável.

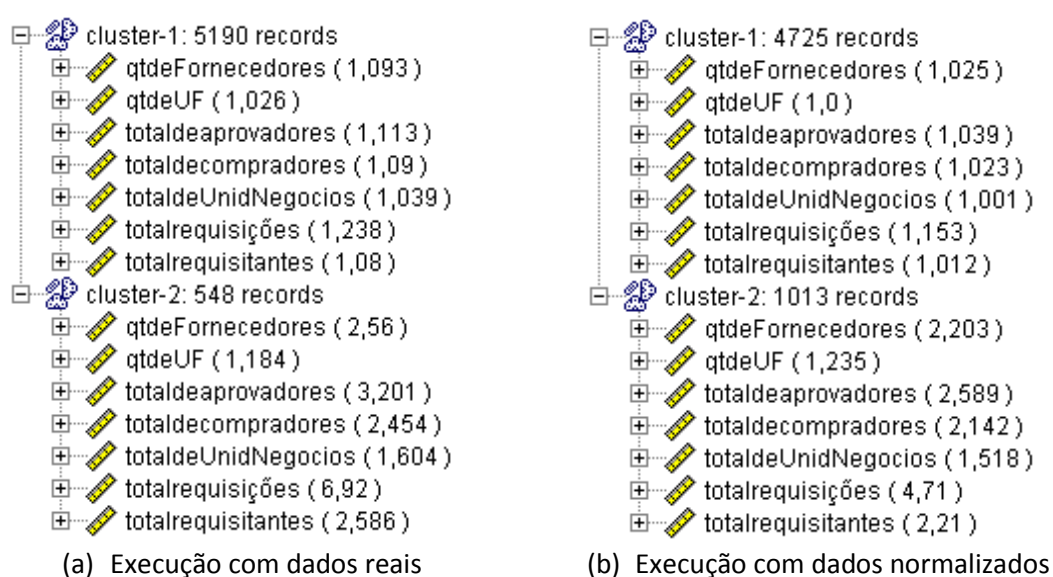


Figura 37 - Two-step cluster – DM MATERIAIS

A seguir, na Tabela 16 e Tabela 17, temos os valores detalhados de cada agrupamento.

Mean	cluster 1		cluster 2	
	real	normalizado	real	normalizado
qtdeFornecedores	1,093	1,025	2,56	2,203
qtdeUF	1,026	1	1,184	1,235
totaldeaprovadores	1,113	1,039	3,201	2,589
totaldecompradores	1,09	1,023	2,454	2,142
totaldeUnidNegocios	1,039	1,001	1,604	1,518
totalrequisições	1,238	1,153	6,92	4,71
totalrequisitantes	1,08	1,012	2,586	2,21

Tabela 16 - Two-step cluster – DM MATERIAIS - Mean

Standard Deviation	cluster 1		cluster 2	
	real	normalizado	real	normalizado
qtdeFornecedores	0,293	0,157	1,31	1,079
qtdeUF	0,16	0	0,406	0,433
totaldeaprovadores	0,321	0,194	1,179	1,121
totaldecompradores	0,29	0,149	0,979	0,851
totaldeUnidNegocios	0,193	0,038	0,839	0,708
totalrequisições	0,521	0,469	5,366	4,616
totalrequisitantes	0,273	0,109	1,169	0,996

Tabela 17 - Two-step cluster – DM MATERIAIS - Standard Deviation.

Tanto usando o recurso de normalização quanto não utilizando, o algoritmo de dois estágios retornou apenas dois clusters. Embora tenhamos diferença na quantidade de indivíduos de cada grupo, a característica do agrupamento em si permaneceu. Identificamos que a maior parte dos registros pertence ao primeiro cluster, não fazendo grande diferença o uso de dados normalizados, conforme podemos observar na Figura 38, abaixo:

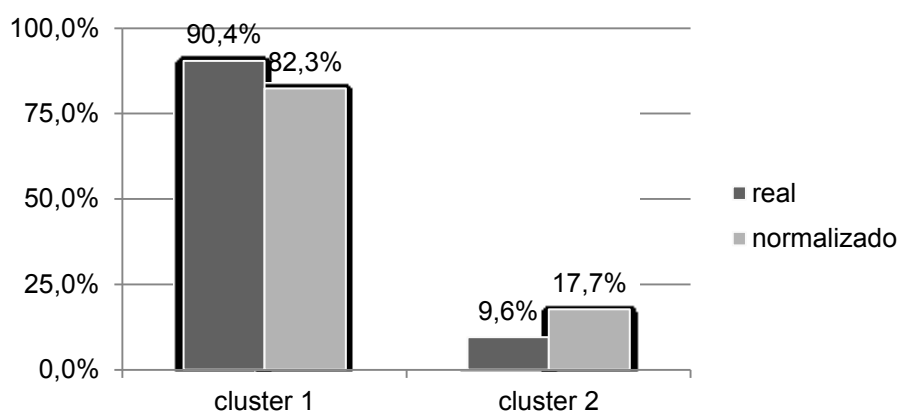


Figura 38 - Two-step cluster – DM MATERIAIS – real x normalizado

Verifica-se que os valores dos centros dos registros no cluster 1 variam entre 1,026 e 1,238. O desvio padrão, valor que quantifica a dispersão dos dados, praticamente não

oscila, com valor mínimo igual a 0,16 e máximo igual a 0,521 – considerando apenas os dados que não foram normalizados. Isto indica uma distribuição normal dos dados.

Para o cluster 2, os valores médios ficam entre 1,184 e 6,92. O desvio padrão apresenta valor mínimo igual a 0,406 e máximo igual a 5,366. Se considerarmos o desvio para os dados normalizados, o valor pouco muda, ficando entre 0,433 e 4,616. Quanto maior o desvio padrão, maior a dispersão e mais afastados da média estarão os eventos mais discrepantes.

As variáveis com maior valor de desvio são *totaldeaprovadores* e *totalrequisições*, mesmo quando considerado a normalização. O que pode indicar que este grupo está associado ao grupo de materiais de uso constante, e utilizados por diversos setores da empresa.

Através da análise da Figura 39 e Figura 40 é possível identificar a influência de cada variável para os agrupamentos. Importante ressaltar que o número de agrupamentos foi estipulado pelo próprio algoritmo de mineração.

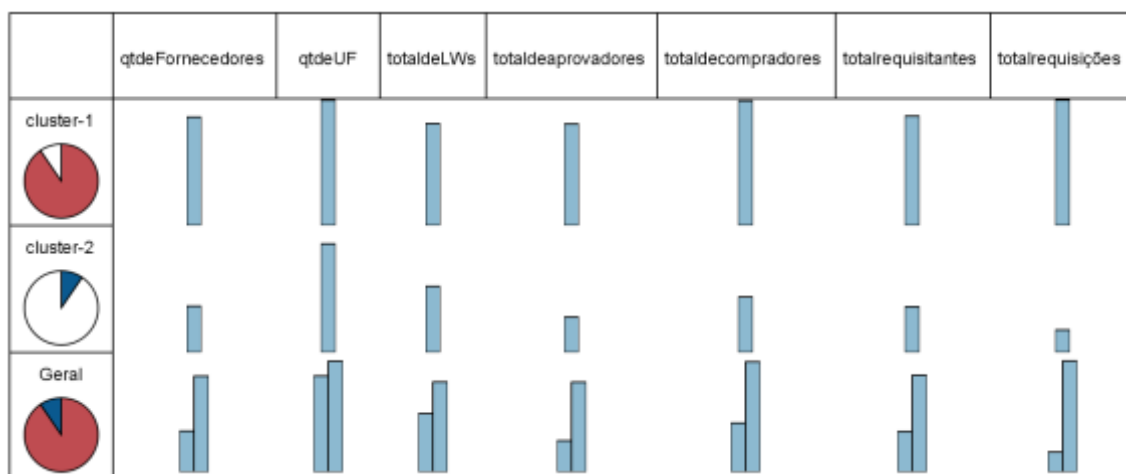


Figura 39 - Two-step cluster – DM MATERIAIS – clusters

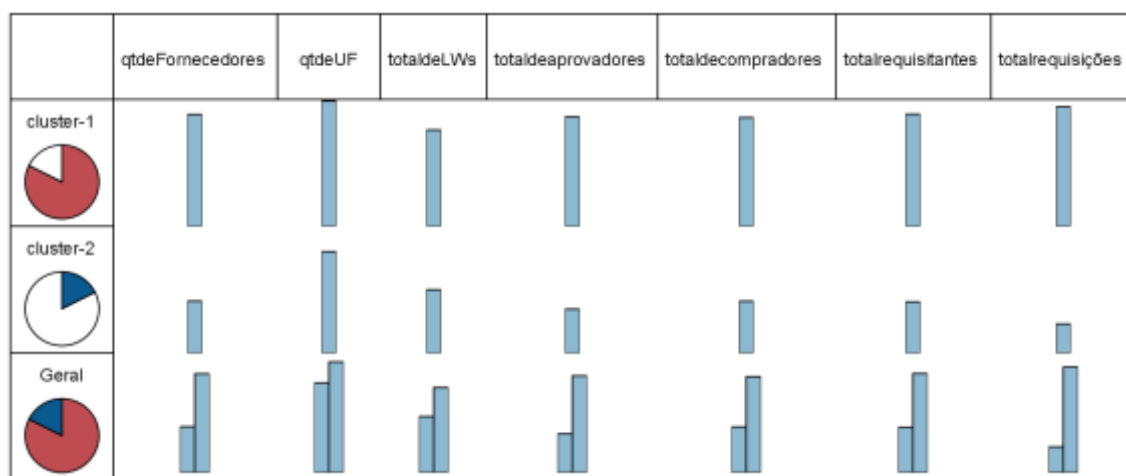


Figura 40 - Two-step cluster – DM MATERIAIS – clusters (normalizado)

A variável *qtdeUF* e *totaldeLWs* (=totaldeUnidNegocios) possuem a mesma distribuição nos agrupamentos. As maiores divergências estão principalmente, segundo

avaliação do especialista, nas variáveis *qtdeFornecedores*, *totalderequisitantes* e *totalrequisições*, que resultará um plano de ação específico para estas observações.

Estes resultados servirão como base de análise para as respostas obtidas com outro clusterizador, o k-means, que veremos a seguir.

### Algoritmo: K-means

Segundo HAN et al. (2001), o k-means é muito popular devido sua facilidade de implementação, mas apesar de sua eficiência, possui a limitação de trabalhar somente com valores numéricos. Difere do algoritmo anterior na questão da definição do número de agrupamentos. Neste, é necessário informar o número de clusters (k) em que os dados deverão ser categorizados. Além deste parâmetro (k), outros dois critérios de parada foram ajustados durante os experimentos:

- ✓ *Maximum Iterations* = 30,0
- ✓ *Change tolerance* = 0,0

Para efeitos comparativos, foram realizados diversos testes para os valores de k. Quando k=2, os grupos obtidos ficaram muito próximos do resultado obtido com o algoritmo two-step cluster – resultados do agrupamento para k=2 e k=3 podem ser vistos na Figura 41.

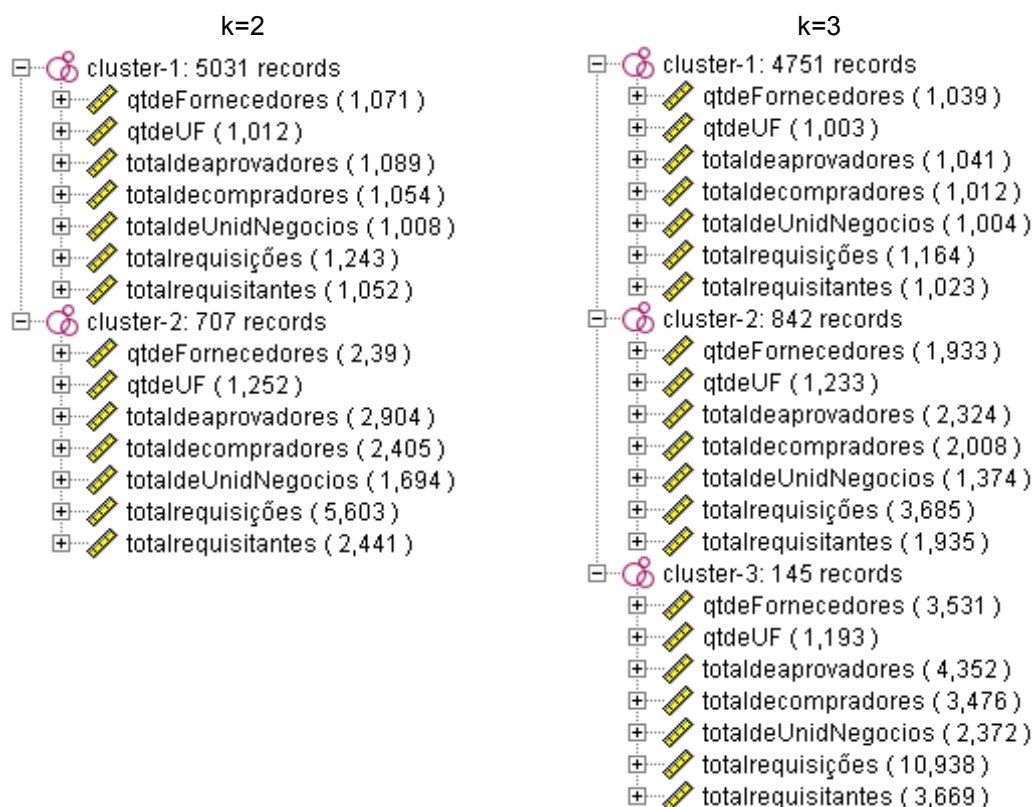


Figura 41 - k-means – DM MATERIAIS – k=2 e k=3

A quantidade de agrupamentos que permitiram análise por parte do usuário especialista varia de três a cinco. Por não utilizarmos variáveis do tipo “set”, o parâmetro de ajuste *Encoding value for sets* não foi modificado.

Os agrupamentos obtidos com os valores de k=4 e k=5 podem ser vistos na Figura 42, que mostra os valores centrais de cada variável.

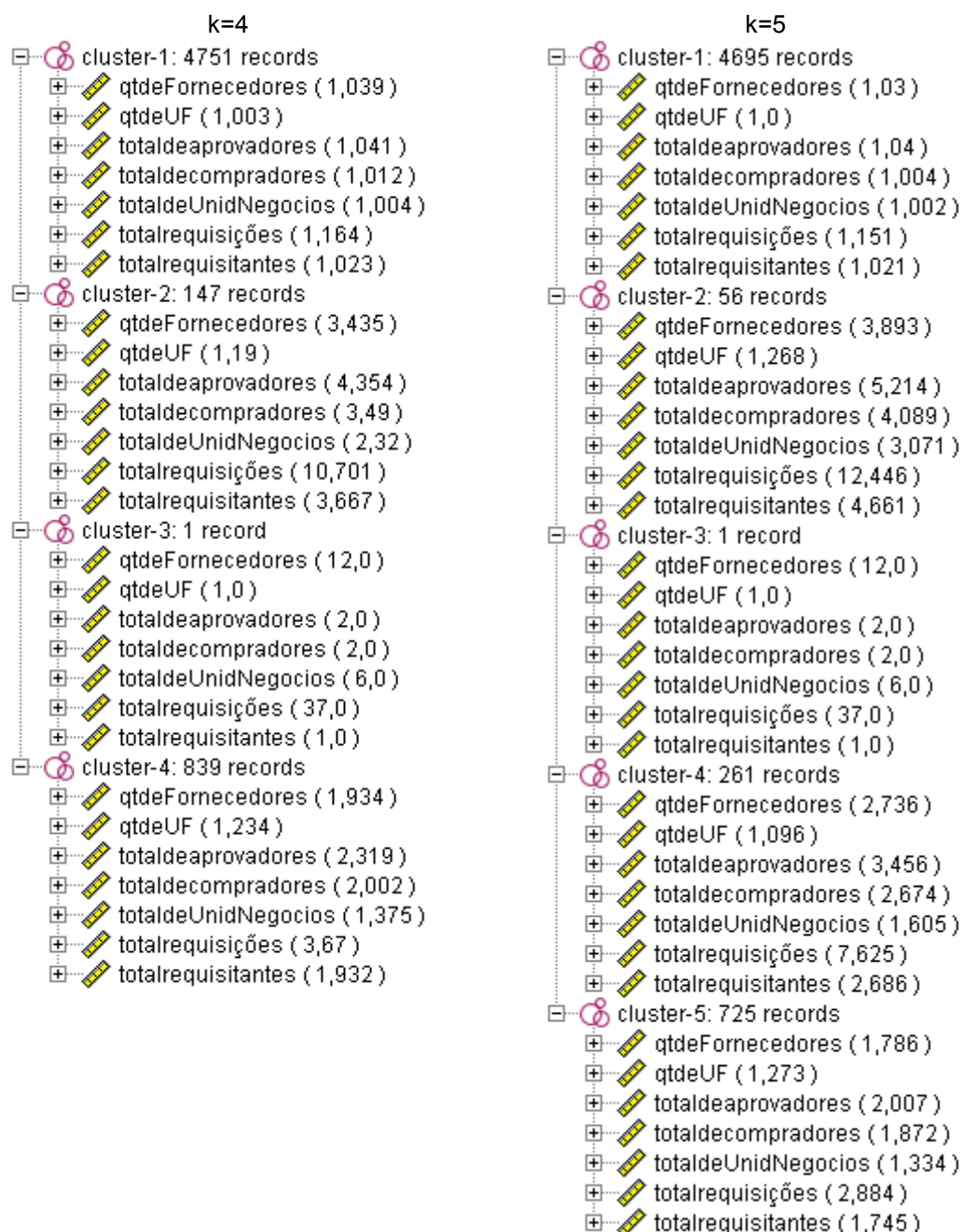


Figura 42 - k-means – DM MATERIAIS – k=4 e k=5

O quadro comparativo com os valores dos centróides de cada agrupamento e seus respectivos desvios encontra-se na Tabela 18 e Tabela 19.

Mean	k=3			k=4			
	cluster 1	cluster 2	cluster 3	cluster 1	cluster 2	cluster 3	cluster 4
qtdeFornecedores	1,039	1,933	3,351	1,039	3,435	12	1,934
qtdeUF	1,003	1,233	1,193	1,003	1,19	1	1,234
totaldeaprovadores	1,041	2,324	4,352	1,041	4,354	2	2,319
totaldecompradores	1,012	2,008	3,476	1,012	3,49	2	2,002
totaldeUnidNegocios	1,004	1,374	2,372	1,004	2,32	6	1,375
totalrequisições	1,164	3,685	10,938	1,164	10,701	37	3,67
totalrequisitantes	1,023	1,935	3,669	1,023	3,667	1	1,932

Standard Deviation	k=3			k=4			
	cluster 1	cluster 2	cluster 3	cluster 1	cluster 2	cluster 3	cluster 4
qtdeFornecedores	0,197	0,802	1,603	0,197	1,453	0	0,802
qtdeUF	0,054	0,431	0,413	0,054	0,411	0	0,432
totaldeaprovadores	0,202	0,75	1,278	0,202	1,259	0	0,746
totaldecompradores	0,108	0,513	0,987	0,108	0,975	0	0,503
totaldeUnidNegocios	0,061	0,515	1,034	0,061	1	0	0,516
totalrequisições	0,512	2,69	7,819	0,512	7,486	0	2,674
totalrequisitantes	0,154	0,663	1,424	0,154	1,406	0	0,662

Tabela 18- k-means – DM MATERIAIS: k=3 e k=4

Mean	k=5				
	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
qtdeFornecedores	1,03	3,893	12	2,736	1,786
qtdeUF	1	1,268	1	1,096	1,273
totaldeaprovadores	1,04	5,214	2	3,456	2,007
totaldecompradores	1,004	4,089	2	2,674	1,872
totaldeUnidNegocios	1,002	3,071	6	1,605	1,334
totalrequisições	1,151	12,446	37	7,625	2,884
totalrequisitantes	1,021	4,661	1	2,686	1,745

Standard Deviation	k=5				
	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
qtdeFornecedores	0,176	1,592	0	1,175	0,637
qtdeUF	0	0,486	0	0,295	0,455
totaldeaprovadores	0,2	1,358	0	0,815	0,577
totaldecompradores	0,067	1,049	0	0,705	0,429
totaldeUnidNegocios	0,048	0,912	0	0,68	0,481
totalrequisições	0,489	8,403	0	5,16	1,643
totalrequisitantes	0,147	1,654	0	0,75	0,587

Tabela 19 - k-means – DM MATERIAIS: k=5

Os resultados obtidos com o k-means na divisão dos dados em 2 agrupamentos ficou muito próximo do encontrado pelo algoritmo Two-step cluster. Além dos valores médios ficarem bastante parecidos, o número de indivíduos em cada cluster pouco variou. Esta diferença pode ser vista na Figura 43, que correlaciona o percentual de indivíduos com o cluster associado.

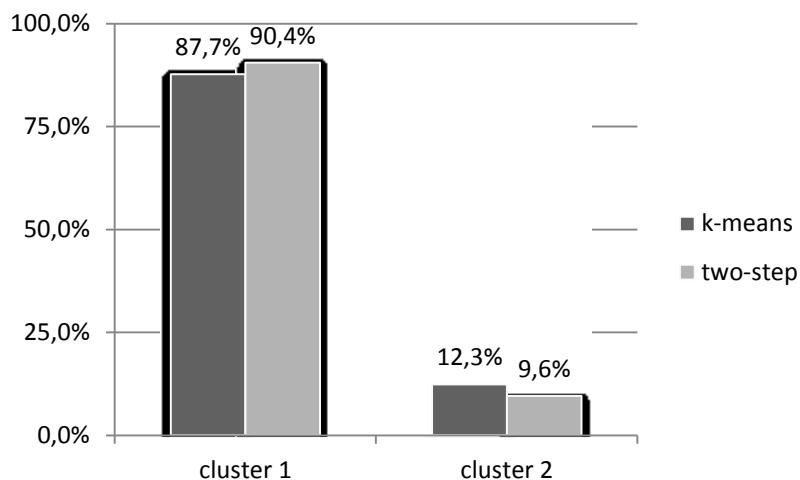


Figura 43 - DM MATERIAIS – comparação entre k-means e two-step para k=2.

Comparando a quantidade de registros para os demais valores de k, Tabela 20, nota-se que para  $k > 3$  temos um tipo de material que destoa dos demais, principalmente para às características relativas a *qtdeFornecedores*, *totaldeUnidNegocios* e *totalrequisições*. Este único indivíduo tornou-se o centro do cluster 3 (para k=4 e k=5).

	k=2	k=3	k=4	k=5
cluster 1	5031	4751	4751	4695
cluster 2	707	842	147	56
cluster 3		145	1	1
cluster 4			839	261
cluster 5				725

Tabela 20 - k-means – DM MATERIAIS: quantidade de registros por agrupamento

Durante a avaliação dos agrupamentos gerados, foram comparados os valores médios das variáveis de cada agrupamento com os valores da Tabela 14. A primeira análise se refere à *qtdeFornecedores* e *qtdeUF*, onde a primeira apresenta valores bastante acima da média em todos os agrupamentos, especialmente no "cluster 3" para k=3 e 4, sendo que nestes grupos o desvio padrão é zero. *QtdeUF* apresenta valores discretos para análise, tendo desvio mínimo igual a 0 e máximo igual a 0,486.

Os agrupamentos encontrados, quando k=3, apresentam características bastante distintas. O primeiro cluster detém aproximadamente 83% dos registros da base de dados e todas as suas variáveis possuem valor de centro próximo a 1, com o desvio padrão menor ou igual a 0,512. De acordo com o usuário especialista, isso demonstra

que a grande parte dos itens comprados atende necessidades extremamente específicas e necessitam de profissionais habilitados para estas negociações.

O segundo cluster, com quase 15% dos registros, apresenta maior variação principalmente no número de requisições. Neste agrupamento nota-se que temos materiais que foram adquiridos por mais de um comprador (considerando os valores médios mais o desvio da variável *totaldecompradores*), mas que dificilmente atenderiam mais de uma unidade de negócios.

O terceiro cluster, ainda para  $k=3$ , detém menos de 3% do total de registros. Caracteriza-se por conter os indivíduos mais solicitados para o departamento de Compras. O especialista do setor de Compras desenvolveu uma análise mais detalhada para este padrão de agrupamento, independentemente do número de registros. A representação gráfica destes agrupamentos está na Figura 44.

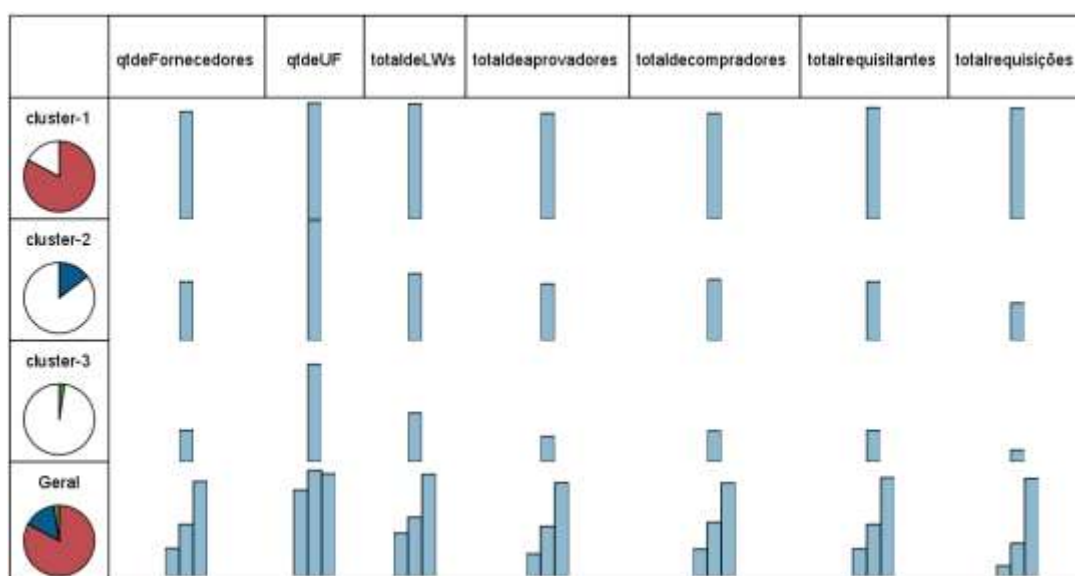


Figura 44 - k-means – DM MATERIAIS:  $k=3$

Para  $k=4$  as características do primeiro cluster não variam, inclusive na quantidade de registros, ficando em torno de 83% do total da base de dados.

No segundo cluster os valores médios de cada variável praticamente dobra, exceto *qtdeUF*. De todos os agrupamentos, este é o que apresenta o segundo maior valor para *totalrequisições*. O segundo cluster representa 3% dos dados. O terceiro agrupamento, conforme dito anteriormente, possui um único indivíduo e representa 0,02% da base.

Para facilitar o entendimento da análise, podemos dizer que o quarto cluster é semelhante ao primeiro, exceto pelos indivíduos com maiores valores para as variáveis *totaldeaprovadores*, *totaldecompradores* e *totalrequisições*, esta última com o maior desvio dentre todas as demais (2,674). O percentual da quantidade de indivíduos para cada agrupamento em função da variação de  $k$  pode ser visto na Tabela 21.

	k=3	k=4	k=5
cluster 1	83%	83%	82%
cluster 2	15%	3%	1%

cluster 3	3%	0,02%	0,02%
cluster 4		15%	5%
cluster 5			13%

Tabela 21 - k-means – DM MATERIAIS: percentual de registros por agrupamento.

O maior valor de k que retornou agrupamentos interessantes para o especialista foi k=5. Para esta quantidade de agrupamentos a maior parte dos registros, aproximadamente 94%, ficou restrita a dois clusters: o cluster 1, com 82% dos registros, e o cluster 5 com 13%. A distribuição de cada variável entre os clusters está na Figura 45.

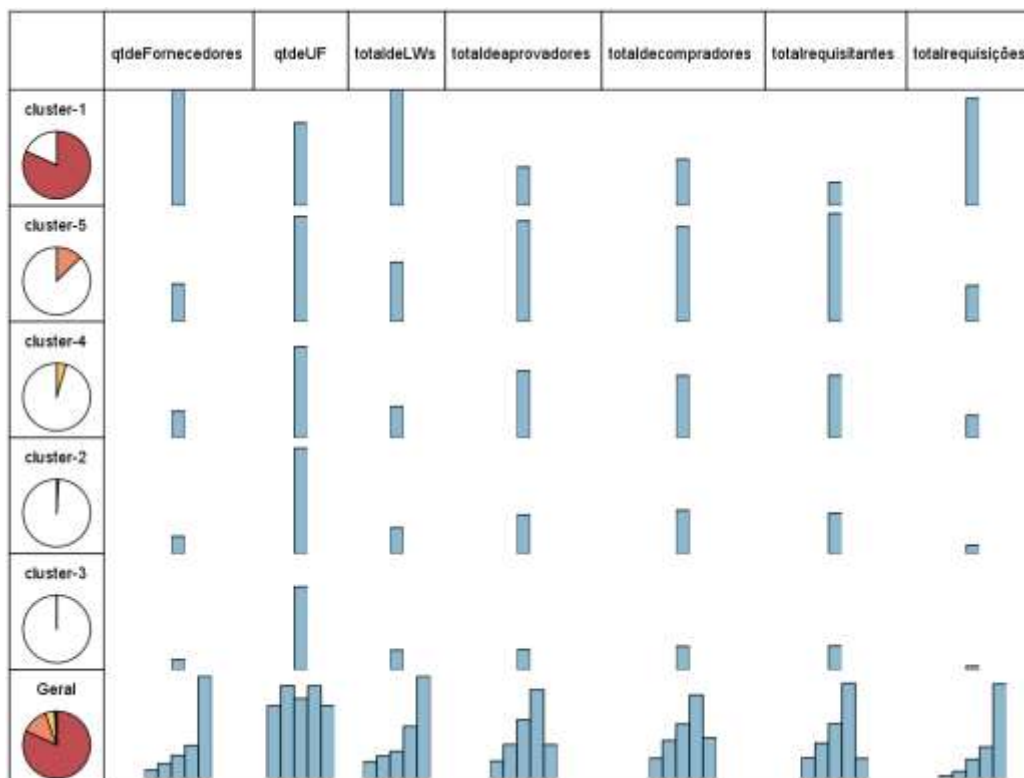


Figura 45 - k-means – DM MATERIAIS: k=3

O terceiro cluster possui um único indivíduo – valor discrepante – e será tratado isoladamente pelo especialista.

O quarto cluster possui alto desvio padrão para a variável *totalrequisições* (5,16) e tem as mesmas características do cluster 2 quando k=4.

A avaliação final do data mart DM MATERIAIS pelo usuário especialista indica o melhor agrupamento obtido para o valor k=3, embora atividades específicas tenham sido desenhadas para o cluster 3 (quando k=4 ou 5).

## Algoritmo: GRI Node

O algoritmo GRI (Generalized Rule Induction) tem como objetivo extrair regras de associação em bases de dados no seguinte formato:

*Se antecedente(s) Então conseqüente(s).*

Para este experimento, foram redefinidos os conjuntos de variáveis utilizadas pelo GRI Node, conforme visto na Tabela 22.

Variáveis da base de dados (data mart)	Variáveis utilizadas pelo algoritmo GRI
codmaterial	codmaterial
totalrequisições	totalrequisições
totalrequisitantes	totalrequisitantes
totaldeaprovadores	totaldeaprovadores
totaldecompradores	totaldecompradores
totaldeUnidNegocios	totaldeUnidNegocios
totaldeCC	qtdeComprada
qtdeNCM	VL_TOTAL_PEDIDO
qtdeComprada	qtdeFornecedores
VL_TOTAL_PEDIDO	qtdeUF
qtdeFornecedores	
qtdeUF	
Total 12	Total 10

Tabela 22- DM Materiais – variáveis utilizadas pelo algoritmo GRI Node.

Em virtude das características da base de dados, um pequeno conjunto de regras foi extraído, mesmo variando os parâmetros de ajuste do algoritmo, como o valor atribuído para o suporte e confiança e o número máximo de antecedentes para cada regra. Na Tabela 23 temos os resultados obtidos após diversas execuções do algoritmo:

Suporte	Confiança	Máximo de antecedentes	Quantidade de regras geradas
0	30%	4	24
0	40%	4	11
0	50%	4	11
0	60%	4	1

Tabela 23 - DM Materiais – parâmetros de ajuste do algoritmo GRI Node.

O resultado que analisaremos a seguir foi obtido após a realização do comando necessário para geração de regras com suporte abaixo de 0.1% e confiabilidade de 30%, a fim de contemplar um grande número de regras de associação.

As regras mais interessantes, de acordo com a avaliação do especialista foram:

id	Consequent	Antecedent	Support %	Confidence %
10	codmaterial 1399009768,000	= totalrequisições > 39,500	0,02	100
11	codmaterial 1399004108,000	= totalrequisitantes > 10,000	0,02	100
12	codmaterial 1399000147,000	= totaldeaprovadores > 7,500 and totalrequisições < 10,500	0,02	100
14	codmaterial 1199000189,000	= qtdeUF > 3,500	0,02	100
15	codmaterial 1325001027,000	= totaldecompradores > 6,500	0,02	100

Tabela 24 - DM Materiais: Regras extraídas pelo algoritmo GRI Node

As regras foram extraídas a partir do conjunto de parâmetros de ajuste - Figura 46:

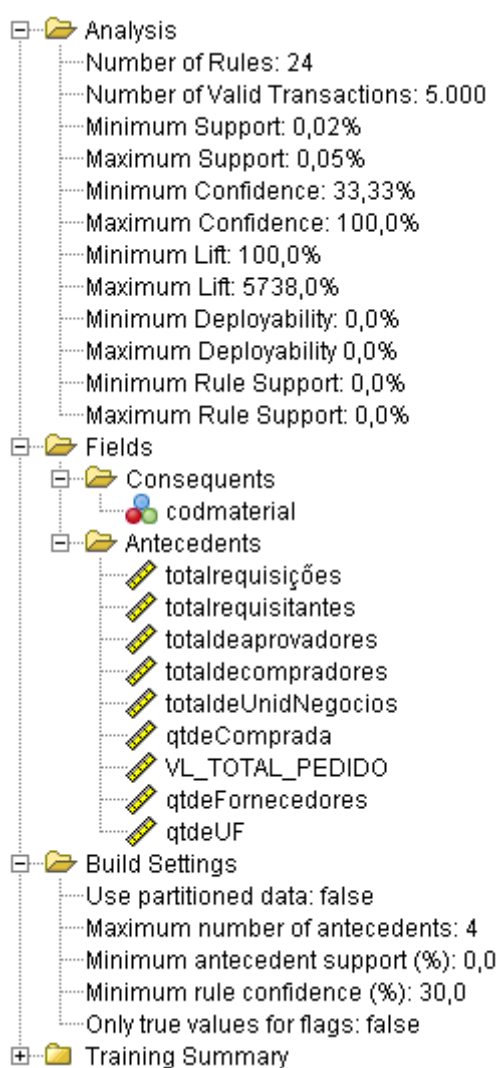


Figura 46- DM Materiais – parâmetros de ajuste do algoritmo GRI Node.

O conjunto de regras obtidas geralmente é de fácil entendimento. Foi concedido autorização para transcrever neste estudo a seguinte tradução das regras obtidas pelo GRI Node:

- ✓ A compra de CADEADO 40 mm foi aprovada por mais de sete aprovadores e possui menos de 11 requisições no período avaliado.
- ✓ O ÓLEO DE CONSERVAÇÃO PARA INOX WURTH 400 ML foi comprado em mais de 3 Estados.
- ✓ Mais de seis compradores compraram OXIGÊNIO.

Segundo o especialista, para a base DM Materiais, os melhores resultados de extração de conhecimento se deram a partir do uso de algoritmos de clusterização.

### **Algoritmo: C5.0 Node**

Para utilizar a técnica de clusterização através da indução de regras de associação é necessário primeiramente analisar as seguintes questões: quantas variáveis serão analisadas conjuntamente e quais serão os níveis de mensuração? O número de variáveis analisadas conjuntamente refere-se a quantas variáveis que se deseja associar numa mesma análise? Qual o resultado esperado? Estes questionamentos levaram o usuário especialista a optar por utilizar o *data mart* DM REQ COMPRAS para avaliar hierarquicamente todo o setor de Compras em função de seus objetivos específicos na organização.

Estes objetivos, conforme discutidos no capítulo 2.1, estão diretamente relacionados com o tipo de envolvimento que o setor possui com seus fornecedores e clientes internos. Por isso, foi utilizado este método de mineração de dados para auxiliar o especialista nas seguintes tarefas:

1. Avaliar, juntamente com o usuário especialista, a estrutura de divisão de trabalho definida para o setor de compras - desenhado na Figura 5. Espera-se obter uma árvore de decisão coerente com a divisão de compradores por unidade de negócios estabelecida.
2. Verificar quais são os prazos mais recorrentes para emissão de um pedido de compra ao fornecedor. Neste caso, procura-se identificar os tempos gastos com negociações e classificar os compradores de acordo com seu desempenho.
3. Relacionar os compradores em função do valor das transações feitas pelo setor. O objetivo é tentar identificar o grupo de pessoas responsáveis pelas negociações críticas.

Em todas as rodadas de testes, foi utilizado os mesmos parâmetros de ajuste no algoritmo C5.0 Node. Utilizamos como resultado a árvore de decisão e o conjunto de regras geradas pelo algoritmo. Os parâmetros estão detalhados na Figura 47.

Após comparar alguns resultados, verificou-se não ser necessário o uso da técnica de reforço (*boosting*), utilizada para melhorar a taxa de acerto do modelo. A opção *Favor*

utilizada foi a padrão do algoritmo, que otimiza a acuracidade, com uma proporção de dados com ruídos ou errados de 5%.

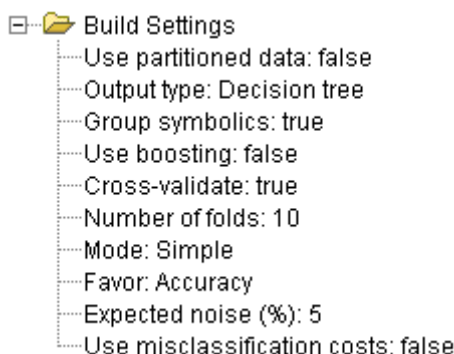


Figura 47 - Parâmetros de ajuste do algoritmo C5.0 Node.

Na avaliação da estrutura de divisão de trabalho definida para o setor de compras foi utilizada apenas duas variáveis categóricas: *UnidNegócio* e *Comprador*. O conjunto de regras obtidas pode ser visto na Figura 48, e a árvore de decisão na Figura 49.

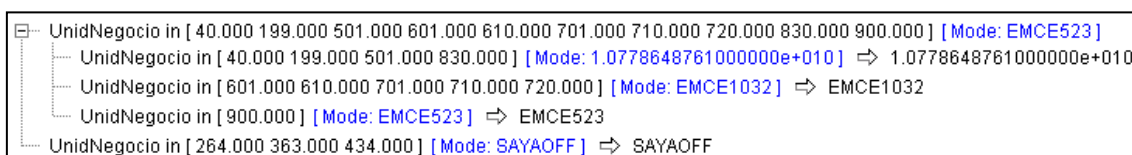


Figura 48 - DM REQ COMPRAS – C5.0 Node - conjunto de regras geradas para as variáveis *UnidNegocio* e *Comprador*.

Como as regras são estruturadas no modelo *SE "antecedência" ENTÃO "consequência"*, podemos ler as regras geradas da seguinte forma:

- SE** as unidades de negócio forem 40, 199, 501, 601, 610, 701, 720, 830 e 900 **ENTÃO** o comprador é EMCE523.
- SE** as unidades de negócio forem 40, 199, 501, e 830 **ENTÃO** o comprador é 07786478610.
- SE** as unidades de negócio forem 601, 610, 701, 710 e 720 **ENTÃO** o comprador é EMCE1032.
- SE** a unidade de negócio for 900 **ENTÃO** o comprador é EMCE523.
- SE** as unidades de negócio são 264, 363 e 434 **ENTÃO** o comprador é SAYAOFF.

Pela leitura das regras o usuário especialista identificou que não existe um agrupamento muito claro, pois diversas linhas de produto (unidades de negócio) são atendidas pelo mesmo comprador, o que contraria a divisão de trabalho definida originalmente.

Quando analisado graficamente através da árvore de decisão da Figura 49, claramente podemos identificar quatro agrupamentos, identificados como nó 2, nó 3, nó 4 e nó 5. O percentual das transações realizadas por cada comprador no respectivo cluster foi a principal informação utilizada pelo especialista neste experimento.

Detalhando os agrupamentos, temos inicialmente o nó 2 (composto pelas unidades de negócio 40, 199, 501 e 830) sendo atendido principalmente pelo comprador 07786487610 (26%), conforme a regra gerada acima. Mas também temos grande

participação de EMCE464 (23%), do comprador 85060679600 (17%) e 52803176400 (14%), além de EMCE523 (11%).

O nó 3 (composto pelas unidades de negócio 601, 610, 701, 710 e 720) tem participação quase que exclusiva de EMCE1032 (66%) e 07786487610 (11%).

O nó 4 (composto pela unidade de negócio 900) tem 65% das transações feitas por EMCE523, as demais transações do nó 4 (35%) foram feitas por praticamente todos os outros compradores em igual quantidade.

Por último, temos o nó 5 (composto pelas unidades de negócio 264, 363 e 434) sendo atendido pelo SAYAOFF (47%), EMCE726 (30%) e EMCE523 (11%).

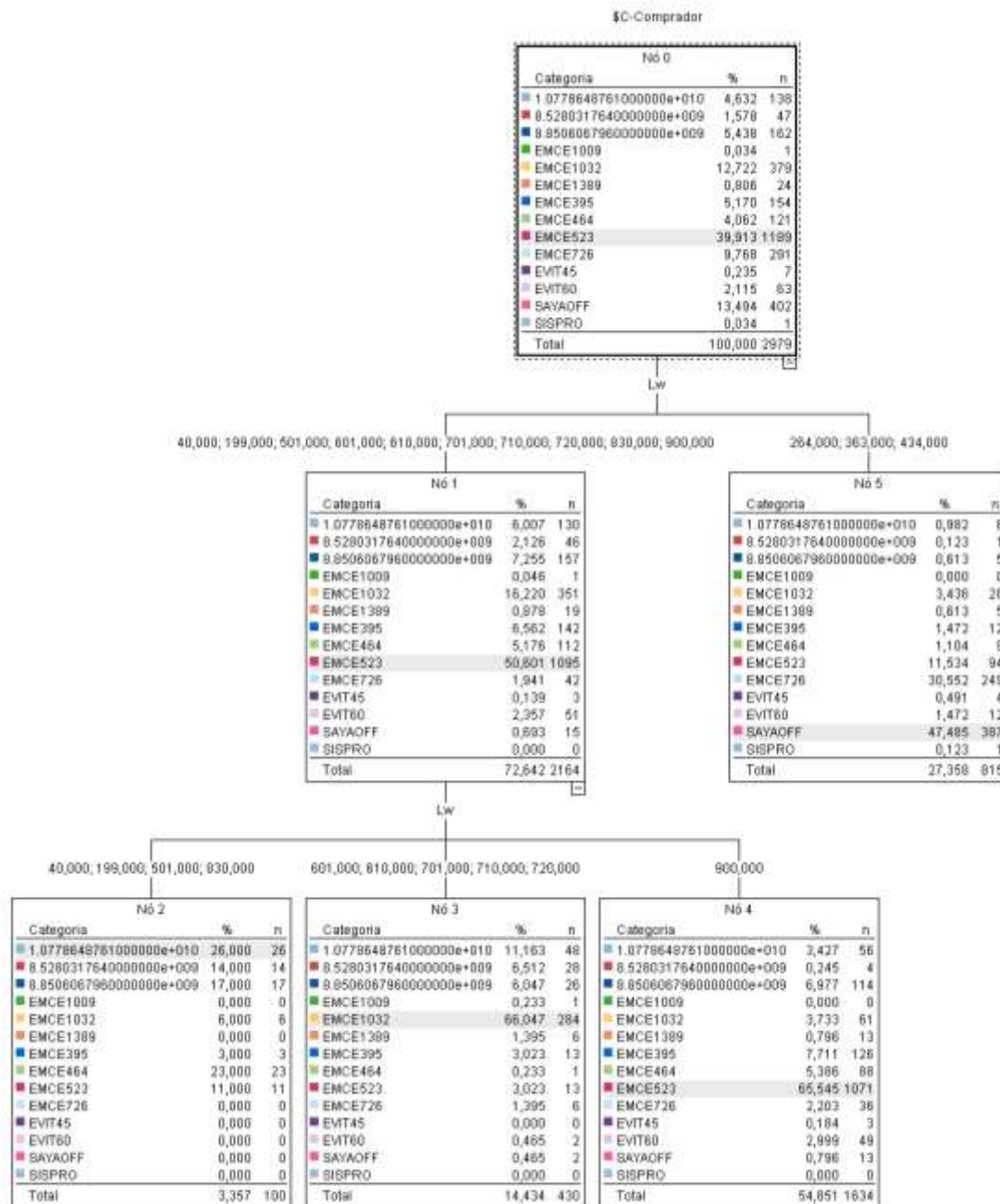


Figura 49 - DM REQ COMPRAS – C5.0 Node - árvore de decisão gerada para as variáveis UnidNegocios e Comprador.

Para verificar os tempos gastos com negociações e classificar os compradores de acordo com seu desempenho utilizamos uma variável numérica, *prazoemissãopedido*, e uma variável categórica, *Comprador*.

O conjunto de regras obtidas neste experimento pode ser visto na Figura 50 e a árvore de decisão na Figura 51. Na figura, podemos a estrutura de regras em apenas dois níveis para facilitar a compreensão e na árvore de decisão, podemos no terceiro nível de detalhamento.

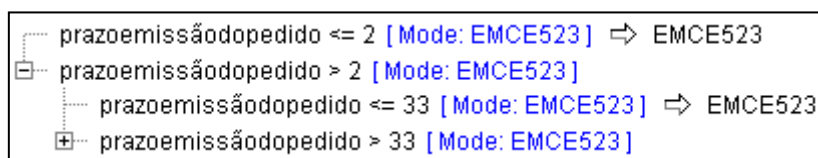


Figura 50 - DM REQ COMPRAS – C5.0 Node - conjunto de regras geradas para as variáveis *Comprador* e *prazoemissãopedido*.

Pelo fato do comprador EMCE523 ter realizado aproximadamente 40% das transações analisadas, as regras em seu primeiro e segundo nível não apresentaram informações úteis, a não ser pelo conjunto de valores relacionados ao prazo de emissão do pedido de compra ao fornecedor.

A árvore de decisão traz informações mais detalhadas, como o nó 1 que representa 23,7% dos dados e agrupa as transações emitidas num prazo menor ou igual a dois dias. Neste nó 1, o comprador EMCE523 e EMCE1032 se destacam com 32,3% e 20% de participação, respectivamente. Importante destacar que todos os compradores estão incluídos em todos os nós da árvore decisão, mas com diferentes níveis de participação.

O segundo bloco de maior importância, de acordo com a avaliação do usuário especialista, é o nó 3. Sua leitura indica que 67% das transações são emitidas com prazo acima de dois e menor ou igual a trinta e três dias. A representação do comprador EMCE523 e SAYAOFF somam 59% neste agrupamento.

O pior caso é representado pelo nó 8, que contém 3,8% das transações analisadas. Este agrupamento inclui os pedidos de compras emitidos com prazo maior que cinquenta e dois dias.

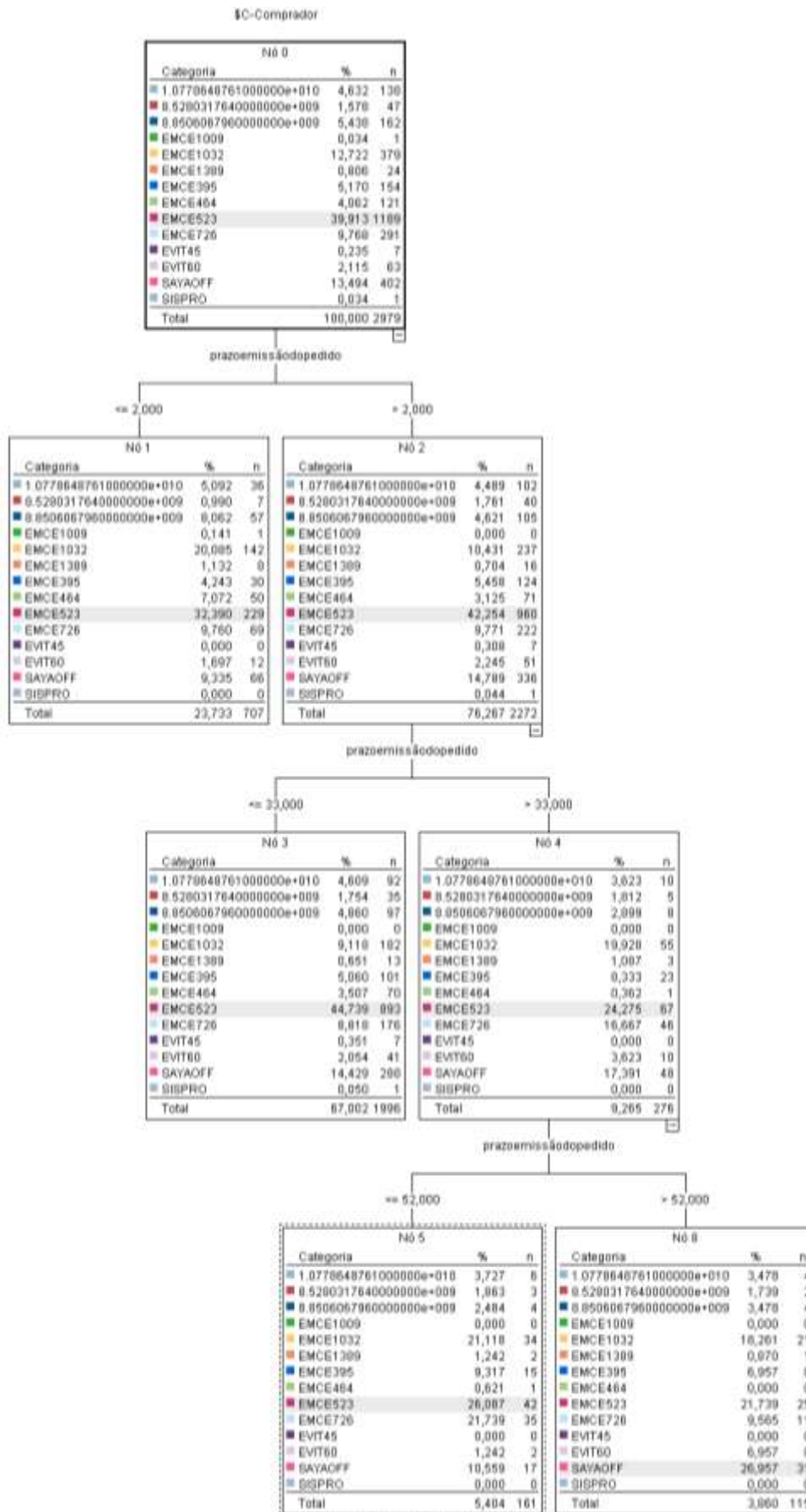


Figura 51 - DM REQ COMPRAS – C5.0 Node - árvore de decisão gerada para as variáveis *Comprador* e *prazoemissãodopedido*.

O último experimento no data mart DM REQ COMPRAS foi feito aplicando o algoritmo C5.0 nas variáveis *comprador* e *soma\_vl\_total\_item*, buscando relacionar os compradores em função do valor das transações feitas pelo setor. O objetivo é tentar identificar o grupo de pessoas responsáveis pelas negociações críticas. O conjunto de regras obtidas após a execução do algoritmo pode ser visto na Figura 52 e a árvore de decisão na Figura 53.

O resultado obtido foi extremamente longo, a árvore de decisão (e conseqüentemente o conjunto de regras) possuía mais de 100 níveis de detalhamento, por isso, foram podadas até apresentarem um resultado satisfatório para o especialista.

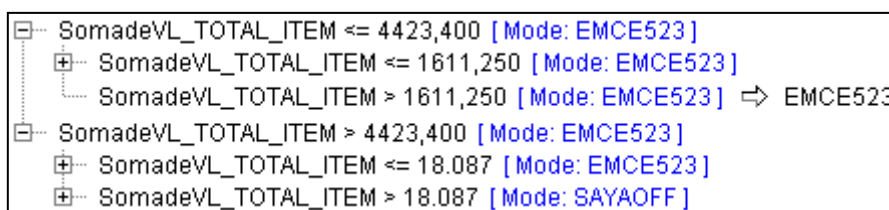


Figura 52 - DM REQ COMPRAS – C5.0 Node - conjunto de regras geradas para as variáveis Comprador e SomadeVL\_TOTAL\_ITEM.

Como acontecido anteriormente, as regras podadas neste caso também não apresentaram informações úteis, pelo fato do comprador EMCE523 ter realizado aproximadamente 40% das transações analisadas. A informação relevante das regras está relacionada aos valores limites de cada grupo, por isso, detalharemos a seguir os pontos mais relevante, observados pelo usuário especialista, em relação à árvore de decisão.

Na árvore temos o nó 1 que representa 76,7% das transações de compras da base de dados. Este agrupamento contém solicitações de compras onde o valor total do pedido, considerando todos os itens, é menor ou igual a R\$ 4.423,40 (quatro mil, quatrocentos e vinte e três reais e quarenta centavos).

O nó 89 (aproximadamente 17% da base de dados) representa todas as compras com valor total do pedido maior que R\$ 4.423,40 e menor que R\$ 18.087,00. O nó 136, com quase 6% dos pedidos do banco de dados, representa todos os pedidos com valor acima de R\$ 18.087,00.

A árvore de decisão deixa visível que todos os compradores estão presentes em todos os agrupamentos, o que indica que não há estrutura hierárquica dos funcionários de compras para negociações estratégicas. Este é outro ponto importante que será tratado pelo usuário especialista.

\$C-Comprador

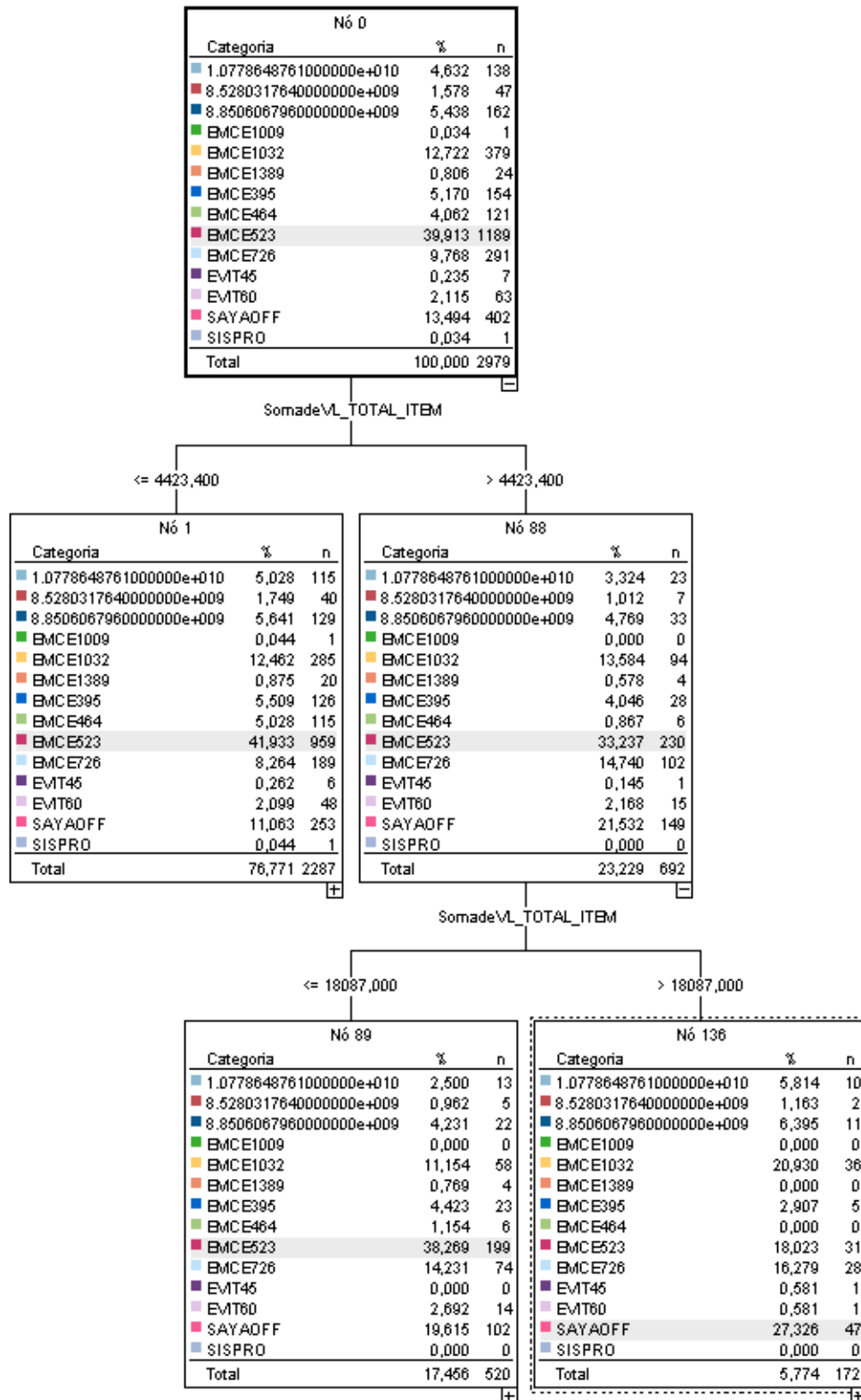


Figura 53 - DM REQ COMPRAS – C5.0 Node - árvore de decisão gerada para as variáveis *Comprador* e *SomadeVL\_TOTAL\_ITEM*.

## 4) Análise dos resultados

O objetivo deste capítulo é descrever a última etapa do processo de Busca de Conhecimento em Banco de Dados – KDD (Knowledge Discovery in Database), no qual segundo STUART et al. (2003), é possível identificar os padrões realmente interessantes entre os diversos obtidos na fase anterior, de acordo com algum critério pré-definido.

Todas as análises foram feitas juntamente com o usuário especialista e serão detalhadas na mesma seqüência em que os algoritmos de mineração de dados foram utilizados. Sempre que possível será demonstrado como estes resultados influenciaram na reestruturação dos mapas estratégicos do setor.

Iniciaremos com o algoritmo **Two-step cluster**, que teve como resposta padrão dois agrupamentos com características bastante distintas independentemente das alterações nas configurações de ajuste do modelo. O primeiro cluster (cluster 1) é composto por aproximadamente 91% das variáveis da base de dados e o segundo, o cluster 2, aproximadamente 9%, conforme pode ser observado na Figura 38.

Pela análise do especialista, os valores de centro das variáveis do cluster 1, Figura 37, representam bem as características dos processos de compras que são feitos no setor, que na maioria das vezes busca atender necessidades específicas de uma operação de uma determinada unidade de negócio. Com isso, é grande a possibilidade da quantidade ter sido estimada de forma incorreta, acarretando em material ocioso no almoxarifado da empresa.

O cluster 2 caracteriza-se pelos materiais de necessidade constante, ou seja, geram ao longo do tempo um grande número de pedidos, de cotações e processos de aprovação. Os materiais do segundo cluster também são utilizados por mais de uma unidade de negócio, o que acaba aumentando a quantidade de solicitantes e, conseqüentemente, de compradores envolvidos no atendimento da compra.

O algoritmo não permitiu uma análise mais precisa, pois a divisão 91% e 9% dificultou bastante a análise, que foi desenvolvida por amostragem de indivíduos em cada cluster, mas serviu como parâmetro inicial de análise.

Para aumentar o número de agrupamentos, foi utilizado outro algoritmo, o **k-means**. Os diversos testes feitos variando o número de clusters (k) resultaram em análises mais refinadas, mas seguiram a mesma linha de raciocínio traçada a partir da análise dos resultados do Two-step cluster.

Conforme pode ser observado nos resultados obtidos com o k-means, Tabela 18 e Tabela 19, a característica e quantidade de indivíduos do primeiro cluster praticamente não variam para os valores de k=2 até k=5. O padrão do primeiro cluster também pode ser reparado nas figuras abaixo (Figura 54 até Figura 57).

A listagem dos materiais que foram agrupados neste primeiro cluster foi encaminhada para o setor de almoxarifado para que seja verificado, juntamente com as unidades de negócio, o que poderia ser reduzido do inventário. A observação incluída no PDCA é *diminuir o inventário e otimizar os processos de compra dos materiais do cluster 1*.

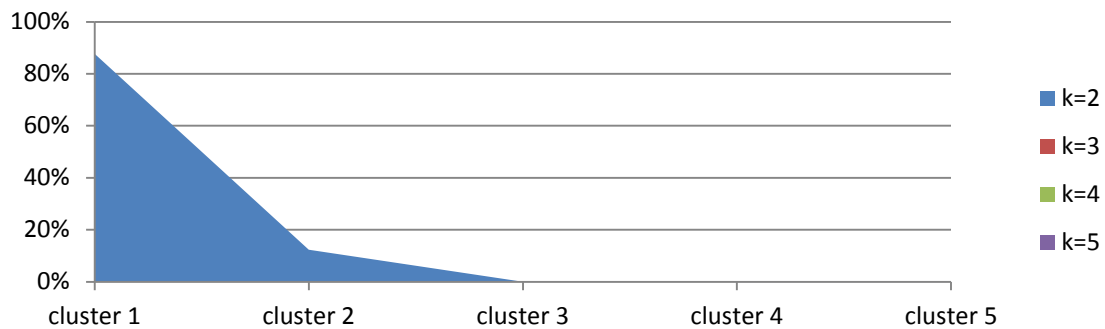


Figura 54 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=2)

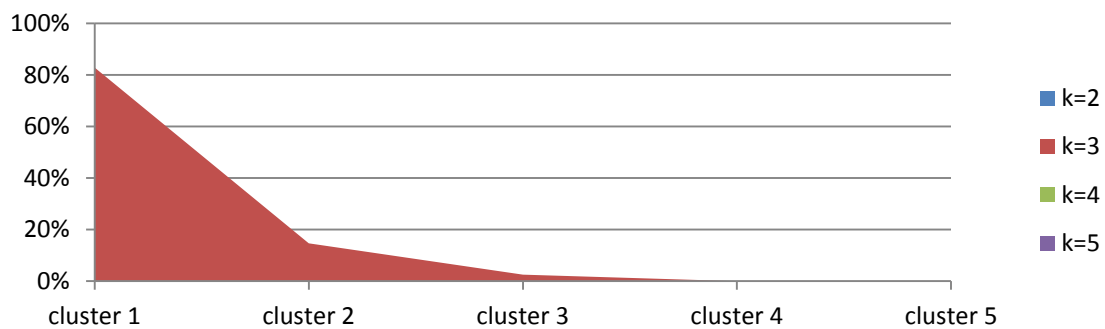


Figura 55 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=3)

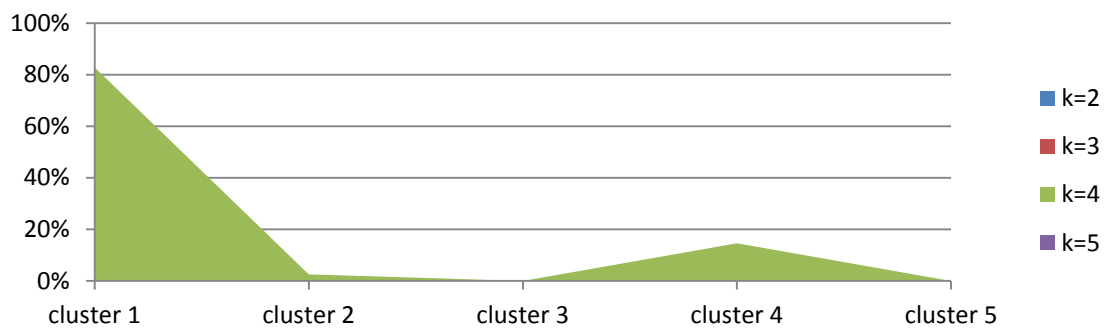


Figura 56 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=4)

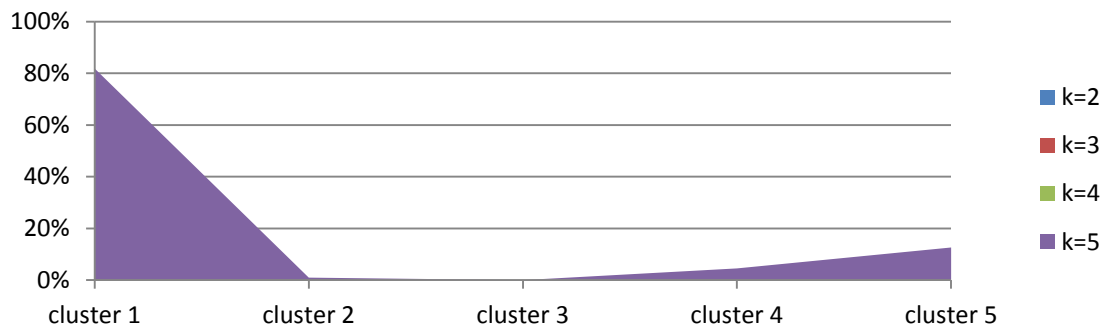


Figura 57 - k-means - Dispersão dos indivíduos da base de dados em função do número de clusters (k=5)

Para o valor de  $k=3$ , o terceiro cluster apresentou características próximas ao segundo cluster quando  $k=2$ . Este agrupamento não foi considerado útil pelo especialista.

O resultado dos agrupamentos obtidos para  $k=4$  e  $k=5$  foi o que trouxe maior credibilidade para esta pesquisa junto à empresa, sua análise foi feita pelo usuário especialista de compras e outro especialista de almoxarifado simultaneamente.

Os extremos dos agrupamentos, ou seja, o primeiro e o último cluster para cada valor de  $k$  são os mais populosos e diferem bastante em sua característica. Pelo perfil de uso e consumo dos materiais que compõem o cluster 5, o cluster 4 e alguns indivíduos do cluster 2, a análise destes agrupamentos levou a empresa a considerar novas tecnologias para aquisição de materiais, como por exemplo o **EDI** (*Electronic Data Interchange*) e o *e-Procurement*.

O **EDI**, conforme visto em ANEFALOS et al. (2001) e GIOVANNINI (2001), é uma ferramenta comercial utilizada na transmissão assíncrona de dados trocados entre clientes e fornecedores, geralmente através da internet. Com o computador da empresa interconectado diretamente ao computador do fornecedor, através de software específico, é possível trocar informações de forma segura sobre pedidos de compra, aviso de recebimento, previsões de entrega, faturamento, cobrança e fluxo de reposição de estoques. Por suas características o EDI é utilizado por empresas com relacionamentos comerciais constantes e com volumes de transações que justifiquem os custos de implantação e manutenção, é o caso, por exemplo, dos materiais contidos principalmente no cluster 4 e no cluster 5.

Segundo COLANGELO (2001), o comércio eletrônico pode ser dividido em duas categorias: o *e-Commerce* (transações de vendas eletrônicas) e o *e-Procurement* (transações de compras eletrônicas). O **e-Procurement** é a ferramenta comercial que faz sistematicamente cotações de preços pela Internet e seleciona fornecedores para participarem de uma concorrência on-line. O software analisa automaticamente as ofertas e escolhe os melhores preços, condições de pagamento e prazo de entrega. Num possível cenário de implementação, as empresas estariam com seus sistemas de venda eletrônica (e-commerce) diretamente conectada aos sistemas de compra eletrônica (e-procurement) da outra empresa. Nesse cenário, seria necessário ter os sistemas integrados também com os sistemas de ERP de cada empresa.

Para estes agrupamentos, duas observações foram incluídas no PDCA da empresa: (i) *realizar análise de aderência para as ferramentas de EDI e e-Procurement* e, (ii) *identificar os fornecedores que já utilizam estas tecnologias*.

Ainda sobre os resultados do k-means, o cluster 3, que pode ser visto na Figura 42, embora classificado pelo especialista como *outlier* por possuir apenas um indivíduo, levantou um importante questionamento: qual o valor gasto com o processo interno de compra na empresa? Esta questão foi colocada após verificar que o valor de mercado do material do terceiro cluster, para  $k=4$  e  $k=5$ , é extremamente insignificante, e que o volume de negócios feitos no período estaria encarecendo o produto devido ao valor

agregado pelo custo<sup>15</sup> do processo interno. Com isso, a seguinte observação foi incluída no PDCA: *buscar fornecedores parceiros para implementar o sistema de loja no almoxarifado da empresa, sendo o armazenamento do produto por conta do fornecedor em área externa à organização.*

Quando analisado os resultados obtidos pelo algoritmo **GRI Node** (Generalized Rule Induction) e **Apriori Node**, observou-se que parte do conhecimento já tinha sido obtida pelas análises anteriores.

O algoritmo **C5.0 Node** permitiu a reavaliação da metodologia de trabalho da equipe de compras. Identificaram-se diversos desvios em relação ao procedimento definido pela empresa e medidas de ajuste foram tomadas.

No procedimento e na estrutura organizacional do setor fica clara a forma como a empresa espera que as unidades de negócio sejam atendidas pelo departamento de compras, onde cada comprador do departamento é responsável pelo atendimento de determinadas linhas de produto, conforme exemplificado na Figura 5 - Divisão de trabalho no setor de compras – elaborada pelo autor. No entanto, os resultados obtidos na aplicação do algoritmo nas variáveis categóricas *UnidNegócio* e *Comprador* (Figura 48 e Figura 49) demonstram que na prática esta organização não vem sendo seguida. Fica claro para o usuário especialista, principalmente após visualizar a árvore de decisão que todos os compradores atendem todas as linhas de produto.

A análise mais detalhada do percentual de participação de cada comprador nas compras das unidades de negócio e o tipo de material comprado levou a inclusão da seguinte observação no PDCA: *alterar o procedimento e a estrutura organizacional do setor de compras de forma que as transações sejam feitas por tipo de material, e não por unidade de negócio.*

O segundo ponto observado pelo especialista refere-se ao tempo gasto para a empresa emitir um pedido de compra ao fornecedor. A árvore de decisão da Figura 51 informa que apenas 23,7% dos pedidos são emitidos num prazo menor ou igual a dois dias. Com esta informação alguns compradores passaram a implementar tabelas de preços padrão, o que melhorou o tempo de resposta do setor. Acredita-se que a modificação da estrutura de atendimento, conforme dito anteriormente, permitirá maior rapidez no processamento das requisições de compras, pois cada comprador ficará focado numa certa quantidade de tipos de materiais. Além da melhoria no prazo, o relacionamento com o fornecedor aumentará, pois ao invés de ter várias pessoas realizando cotações e negociando preços isoladamente, teremos apenas um ponto de contato.

Finalmente, no último experimento feito neste estudo foi avaliada a relação entre o valor das transações feitas pelo setor e as pessoas responsáveis pelas negociações. Assim, foi aplicado o algoritmo **C5.0 Node** na variável categórica *comprador* e na variável numérica *soma\_vl\_total\_item*.

Na

---

<sup>15</sup> Aqui foi incluído o valor/hora de cada departamento para: registrar a requisição no sistema, analisar a requisição e fazer as devidas cotações, consolidar o menor preço, buscar autorização financeira junto à gerência e, por fim, emitir o pedido de compra.

Figura 52 e Figura 53 é possível verificar que o conjunto de regras geradas e a árvore de decisão possuíam mais de 100 níveis de detalhamento e foram podadas até apresentarem um resultado satisfatório para o especialista. O fato do comprador EMCE523 ter realizado aproximadamente 40% de todas as transações realizadas no período acarretou em regras genéricas, que na maioria das vezes apresentaram como consequência o mesmo comprador.

O conhecimento extraído neste caso está relacionado aos grupos de valores negociados por requisição, que segundo o especialista, podem ser classificados de acordo com a Tabela 25:

	Pedidos	%	Parâmetro			
grupo 1	2.287	76,77%	Entre	R\$ 0,01	e	R\$ 4.423,40
grupo 2	520	17,46%	Entre	R\$ 4.423,41	e	R\$ 18.087,00
grupo 3	172	5,77%	Acima de	R\$ 18.087,01	até	R\$ 1.022.129,34

Tabela 25- Grupos de valores negociados por requisição

Os três grupos de valores de compras identificados acima serão compostos por compradores específicos que receberão treinamento específico sobre Técnicas de Negociação. Esta é foi a última observação incluída pelo usuário especialista no PDCA: *definir o grupo de compradores responsáveis por cada faixa de compra definida na Tabela 25 e agendar os treinamentos necessários.*

## 5) Conclusão

Neste trabalho foi apresentado um estudo descritivo exploratório realizado em uma empresa multinacional do setor de petróleo e gás que possui base operacional na cidade de Macaé, estado do Rio de Janeiro.

A revisão bibliográfica permitiu melhor entendimento sobre a aplicação da estratégia aprendizacional defendida por BARCELLOS (2004), que correlaciona de forma sistêmica o uso de sistemas inteligentes (como as ferramentas de mineração de dados) com o planejamento estratégico das organizações. Este relacionamento se deu nesta pesquisa através da aplicação de diversas técnicas de extração de conhecimento na base de dados do módulo de suprimentos do sistema ERP corporativo.

A construção dos mapas de causa e efeito dos objetivos estratégicos foi iniciada depois de discutido com os especialistas do departamento o estudo feito sobre a evolução do setor de compras, apresentado no capítulo 2.1. A discussão permitiu o alinhamento da metodologia de análise e melhoria de processos que foi utilizada, bem como os resultados esperados pela empresa.

Para a realização dos experimentos foi configurado um ambiente de desenvolvimento com uma cópia das transações de compras realizadas entre 01 de janeiro de 2008 e 31 de dezembro de 2009. Para a manipulação dos dados foi desenvolvido em Delphi uma ferramenta de *Business Intelligence* baseada no componente OLAP *PivotCube*, mantido pela empresa PivotWareLab. Esta ferramenta foi utilizada no ambiente de desenvolvimento nas etapas de seleção de atributos e pré-processamento. Ao término do estudo esta ferramenta foi migrada para o ambiente de produção e até o momento é utilizada diariamente pelos funcionários da empresa.

Durante os testes feitos com os algoritmos de mineração de dados, ficou clara a necessidade do acompanhamento do especialista nas avaliações dos resultados para dar suporte ao processo de tomada de decisão. Pela característica dos dados, os melhores resultados de mineração foram obtidos a partir das técnicas de clusterização e árvores de decisão. As regras de associação não apresentaram informações relevantes para o usuário especialista.

Após o término da fase experimental, detalhada no capítulo 3, o conhecimento extraído resultou em novas propostas de uso de tecnologia e diversos planos de ação para a reestruturação da metodologia de trabalho do setor de compras da empresa, conforme enumeramos abaixo:

1. Foi comprovada a necessidade de reestruturação do setor em função da distribuição das compras, seja por tipo de material ou valor da compra. Inicialmente, por recomendação do especialista, foi incluída no PDCA da corporação a necessidade de alterar o procedimento e a estrutura organizacional do departamento de forma que as transações sejam processadas por tipo de material, e não por unidade de negócio como é feito atualmente.
2. Foi identificada uma grande oportunidade de redução de custos através do direcionamento de compradores específicos para cada um dos três grupos de valores negociados por requisição, apresentado na Tabela 25. Para esta ação, foi incluída no

PDCA da empresa a necessidade de treinamentos específicos para os responsáveis por cada grupo.

3. Os resultados apresentados pelo algoritmo *k-means* quando o número de agrupamentos é definido em  $k=4$  ou  $k=5$ , sugerem o uso de novas tecnologias aplicadas ao processo de compras, como o *EDI (Electronic Data Interchange)* e o *e-Procurement*. O usuário especialista incluiu para esta avaliação os seguintes planos de ação para o setor:
  - (i) Buscar fornecedores para realizar análise de aderência para as ferramentas de *EDI* e *e-Procurement* junto ao ERP Sispro.
  - (ii) Identificar os fornecedores que já utilizam estas tecnologias.
4. As características do primeiro cluster obtido em todas as variações de “ $k$ ” do algoritmo *k-means* indicaram a oportunidade de redução de inventário para estes itens. O estudo apresentou ao departamento de almoxarifado uma listagem com o código de todos os materiais que deverão ser analisados.
5. Foi constatado que o custo associado ao processo de compra (incluindo o valor/hora dos setores para concretizar uma compra) estaria encarecendo alguns materiais identificados pela análise de agrupamentos, devido ao seu baixo valor de mercado. Esta observação levou a empresa a buscar fornecedores parceiros para implementar o *sistema de loja* no almoxarifado da empresa, sendo o armazenamento do produto por conta do fornecedor em área externa à organização.

Nesta pesquisa foram avaliadas as transações de compras de materiais sem levar em consideração os dados relacionados ao consumo (retirada do estoque), o que certamente limitou as análises por não permitir comparações *fim-a-fim* em toda a cadeia de suprimento. Por isso, como trabalho futuro, o estudo sugere que novas bases de dados específicas sejam criadas incluindo informações de outras tabelas do sistema ERP, como por exemplo, as tabelas de movimentação de estoque, solicitações de pedidos, informações do ativo fixo e o custo total de propriedade (ou TCO do inglês: *total cost of ownership*) do bem comprado.

Outra oportunidade de trabalho futuro seria a integração dos modelos gerados pelos algoritmos de mineração de dados com a ferramenta OLAP baseada no componente *PivotCube*.

Por último, é importante ressaltar que os resultados da mineração foram obtidos a partir de dados históricos. Por isso, é necessário manter uma regularidade na avaliação dos modelos e agrupamentos obtidos. O presente estudo sugere que anualmente seja feito o processo de mineração e análise para identificar as melhorias e, assim, progredir com a rampa de melhoria de processos, demonstrada na Figura 8.

## 6) Referências Bibliográficas

- AGRAWAL, RAKESH; IMIELINSKI, TOMASZ; SWAMI, ARUN. **Mining Association Rules between Set of Items in Large Databases**. In: ACM sigmod int'l conference on management of data, 1993.
- AGRAWAL, RAKESH; SRIKANT, RAMAKRISHNAN. **Fast Algorithms for Mining Association Rules**. In: 20th int'l conference on very large databases, 1994.
- ANDRADE, FÁBIO F. **O método de melhorias PDCA**. Dissertação de Mestrado. Escola Politécnica Universidade de São Paulo. São Paulo - SP, 2003.
- ANEFALOS, LILIAN C; CAIXETA FILHO, JOSÉ V. **Tecnologia de Informação e Sua Influência Sobre os Rumos da Comercialização de Produtos**. Revista Informação & Informação, v. 6. Londrina, 2001. Disponível em <<http://www.uel.br/revistas/informacao/include/getdoc.php?id=304&article=101&mode=pdf>>. Acessado em 13/01/2011.
- BARCELLOS, PAULO CESAR DE ARAUJO. **Estratégia aprendizacional: integração dos conceitos de Balanced Scorecard, comunidades virtuais e sistemas inteligentes**. Tese Doutorado. Universidade Federal do Rio de Janeiro - RJ, 2004.
- BETHLEM, A S. **Avaliação ambiental e competitiva**. 3ª edição. Rio de Janeiro. AGIR, 1996.
- BISHOP, CHRISTOPHER M. **Pattern Recognition and Machine Learning**. Editora Springer, 2006.
- BRAGA, ATAÍDE. **Evolução estratégica do processo de compras ou suprimentos de bens e serviços nas empresas**. Disponível em <[http://www.centrodelogistica.com.br/new/art\\_Evol\\_Estrat\\_de\\_compras\\_e\\_supr\\_bens\\_de\\_serv.pdf](http://www.centrodelogistica.com.br/new/art_Evol_Estrat_de_compras_e_supr_bens_de_serv.pdf)> acessado no dia 06 de janeiro de 2010.
- CALINSKI, T.; HARABASZ, J. **A Dendrite Method for Cluster Analysis**. Communications in statistic, v.3,pp.1-27, 1974.
- CAMPOS, VICENTE FALCONI. **Controle da Qualidade Total (no estilo Japonês)**. Fundação Christiano Ottoni - Belo Horizonte, 1992.
- CAMPOS, VICENTE FALCONI. **Gerenciamento da rotina do trabalho do dia-a-dia**. Editora Desenvolvimento Gerencial. Belo Horizonte, 2001.
- CANIËLS, MARJOLEIN C. J.; GELDERMAN, CEES J. **Purchasing strategies in the Kraljic matrix — A power and dependence perspective**. Journal of Purchasing & Supply Management 11, 2005.
- CARLANTONIO, LANDO MENDONÇA DI. **Novas Metodologias para Clusterização de Dados**. Tese Mestrado – COPPE, UFRJ. Rio de Janeiro, 2001.

- CARNEIRO, TERESA CRISTINA JANES. **Integração organizacional e tecnologia da informação: um estudo na indústria farmacêutica**. Tese Doutorado, UFRJ – COPPEAD. -Rio de Janeiro, 2005.
- CHRISTOPHER, M. **Logística e gerenciamento da cadeia de suprimentos**. São Paulo. Pioneira, 1997.
- COLANGELO FILHO, LUCIO. **Implantação de sistemas ERP: um enfoque de longo prazo**. São Paulo: Atlas, 2001.
- CORRÊA, HENRIQUE L.; GIANESI, IRINEU G. N.; CAON, MAURO. **Planejamento, Programação e Controle da Produção. MRP II / ERP - Conceitos, Uso e Implantação**. 4ª Edição, Atlas, 2001.
- DJAIR. Gráfico de Pareto. Disponível em <http://www.lugli.org/2008/02/grafico-de-pareto/>, acessado em 22 de janeiro de 2010 às 17 horas. Lugli.org, 2008.
- FERRAZ, J.C; KUPFER, B; HAUGUENAUER, L. **Made in Brazil: desafios competitivos para a indústria**. Rio de Janeiro: Campus, 1997.
- GIL, ANTONIO CARLOS. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 1991.
- GIOVANNINI, FABRIZIO. **A Empresa Média Industrial e a Internet**. Caderno de Pesquisas em Administração, USP, V. 8, São Paulo, 2001. Disponível em: <<http://www.ead.fea.usp.br/Cad-pesq/arquivos/v8-3-art01.pdf>>. Acesso em 12/01/2011.
- GOLDRATT, ELIYAHU M.; COX, JEFF. **A meta: um processo de aprimoramento contínuo**. Editora Educator, São Paulo, 1997.
- HAN, J.; KAMBER, M. **Data Mining – Concepts and Techniques**. 1ª edição. Nova York: Morgan Kaufmann, 2001.
- HAVE, STEVEN TEM; HAVE, WOUTER TEM; STEVENS, FRANS; ELST, MARCEL VAN DER. **Modelos de gestão: o que são e quando devem ser usados**. São Paulo: Prentice Hall, 2003.
- KAPLAN, ROBERT. S.; NORTON DAVID P. **Mapas Estratégicos - Balanced Scorecard: convertendo ativos intangíveis em resultados tangíveis**. Editora Elsevier, Rio de Janeiro – RJ, 2004.
- KOHAVI R; JOHN GH. **Wrappers for feature subset selection**. Artificial Intelligence 97, pp. 273-324, 1997.
- KUME, H. **Métodos estatísticos para melhoria da qualidade**. Ed. Gente, 9ª edição. São Paulo, 1993.

- LAPOINTE, F. J.; LEGENDRE, P. **The Generation of Random Ultrametric Matrices Representing Dendrograms**, Journal of Classification, 8, pp. 177-200, 1991.
- LAUDON, KENNETH C.; LAUDON, JANE P. **Sistemas de informação gerenciais: administrando a empresa digital**. 5. ed. Prentice Hall. São Paulo - SP, 2004.
- MARKOV, ZDRAVKO, LAROSE, DANIEL T. **Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage**. Wiley-Interscience, 2007.
- MEYER, ANDRÉIA DA SILVA. **Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes**. Tese de Mestrado - Piracicaba, 2002.
- MONCZKA, R. M, TRENT, R. J. **Purchasing and supply management: trends and changes throughout the 1990s**. International Journal of Purchasing and Materials Management vol. 34, 1998.
- MONCZKA, R.; TRENT, R.; HANDFIELD, R. **Purchasing and Supply Chain Management**. Thompson, 2003.
- MONTEIRO, ANDRÉ V. G.; PINTO, MARCOS PAULO O.; COSTA, ROSA MARIA E. M. **Uma aplicação de Data Warehouse para apoiar negócios**. Cadernos do IME - Instituto Militar de Engenharia. Série Informática, vol. 16, Junho de 2004.
- MOORI, ROBERTO GIRO, MARCONDEZ, REYNALDO CAVALHEIRO, ÁVILA, RICARDO TEIXEIRA. **A análise de agrupamentos como instrumento de apoio à melhoria da qualidade dos serviços aos clientes**. Artigo publicado na RAE, Vol. 6. Rio de Janeiro, 2002.
- MOURA, REINALDO APARECIDO; BANZATO, JOSÉ MAURÍCIO. **Jeito Inteligente de Trabalhar: 'Just-in-Time' a reengenharia dos processos de fabricação**. São Paulo: IMAM, 1994.
- NEVES, THIAGO FRANCA. **Importância da utilização do Ciclo PDCA para garantia da qualidade do produto em uma indústria automobilística - EPD/UFJF**, 2007.
- OLIVEIRA, DJALMA. DE P. R. **Planejamento estratégico: conceitos, metodologia e prática**. 15ª ed. São Paulo: Atlas, 2001.
- OLIVEIRA, DJALMA. DE P. R. **Sistemas, Organização & Métodos**. São Paulo: Atlas, 2002.
- PAKHIRA, M. K., BANDYOPADHYAY, S., MAULIK, U. **Validity index for crisp and fuzzy clusters**, Pattern Recognition, June 2004.
- PEARSON, J. N. **A longitudinal study of the role of the Purchasing function: toward team participation**. European Journal of Purchasing & Supply Management. Vol. 5, 1999.

- PORTER, MICHAEL E. **Vantagem competitiva: criando e sustentando um desempenho superior**. Rio de Janeiro: Campus, 1990.
- PROTIL, ROBERTO MAX; SOUZA, VINICIUS PISSAIA. **Sistemas de informação e cadeia de suprimentos**. XXVI ENEGEP - Fortaleza, 2006.
- QUINLAN, J. R. **Induction of decision trees**. Machine Learning 1, 81-106. 1986.
- QUINLAN, J. R. **C4.5: Programs for Machine Learning**. Morgan Kaufmann, San Francisco, 1993.
- RADDING, ALAN. Artigo: **"It's in the can: Analytical applications simplify back-end datamarts"**, in "Datamation", 1999.
- SANTOS, A. B.; MARTINS, M. F. **Medição de desempenho e alinhamento estratégico: requisitos para o sucesso do Seis Sigma**. In. Simpósio de Administração da Produção, Logística e Operações Internacionais. Anais São Paulo, SP, 2005.
- SENGE, PETER. **A quinta disciplina: Arte e prática nas organizações**. São Paulo: Best Seller, 1998.
- SOUZA, CESAR ALEXANDRE DE. **Sistemas integrados de gestão empresarial: estudos de caso de implementação de sistemas ERP**. São Paulo, FEA/ USP, 2000.
- SOUZA, R. de. **Metodologia para desenvolvimento e implantação de sistemas de gestão da qualidade em empresas construtoras de pequeno e médio porte**. Tese Doutorado. Escola Politécnica Universidade de São Paulo, 1997.
- STUART RUSSEL; PETER NORVIG. **Artificial Intelligence, a modern Approach**. Second edition, Prentice Hall, 2003.
- TYRON, R.C. **Cluster Analysis**. Ann Arbor, MI: Edwards Brothers, pp. 422, 1939.
- VIEIRA, S. **Estatística para a qualidade: como avaliar com precisão a qualidade em produtos e serviços**. Rio de Janeiro: Campus, 1999.
- WITTEN, I.H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**. Morgan Kaufmann Publishers, 2nd edition, San Francisco, California, 2005.

# Anexo I – Comandos SQL

## Stored Procedure RELATORIO\_ENTRADAS

```
set ANSI_NULLS ON
set QUOTED_IDENTIFIER ON
go

ALTER procedure [dbo].[RELATORIO_ENTRADAS] @estab varchar(4), @lw varchar(50),
@requisitante varchar(30), @comprador varchar(30), @codforn varchar(30), @nomeforn
varchar(30), @semDoc varchar(2), @por varchar(2), @documento varchar(10), @dtini
datetime, @dtfim datetime AS

declare @sql as varchar(8000)

set @sql = ' select * ,'''+@POR+'''' DOCUMENTO'
if @por = 'RC'
    set @sql = @sql + ', YEAR(dt_rqco) AS ANO, MONTH(dt_rqco) AS MES'
else if @por = 'PC'
    set @sql = @sql + ', YEAR(dt_pdco) AS ANO, MONTH(dt_pdco) AS MES'
else
    set @sql = @sql + ', YEAR(dt_nent_entrada) AS ANO, MONTH(dt_nent_entrada) AS MES'

set @sql = @sql + ' from [bra-sql-rj].sispro_2008v26.dbo.BI_entradas where
cd_estab='''+@estab+''''
if @lw<>''
    set @sql = @sql + ' and (LW LIKE ''%'+@LW+'%' OR '''+@LW+'''' LIKE ''%'+LW+'%'')'
if @requisitante<>''
    set @sql = @sql + ' and (nomeuserrqco LIKE ''%'+@requisitante+'%' OR
'''+@requisitante+'''' LIKE ''%'+nomeuserrqco+'%'')'
if @lw<>''
    set @sql = @sql + ' and (comprador LIKE ''%'+@comprador+'%' OR '''+@comprador+''''
LIKE ''%'+comprador+'%'')'
if @codforn<>''
    set @sql = @sql + ' and isnull(cd_pessoa_forn, cd_pessoa_remet)=''''+@codforn+''''
if @nomeforn<>''
    set @sql = @sql + ' and (isnull(fornecedor, razao) LIKE ''%'+@nomeforn+'%' OR
'''+@nomeforn+'''' LIKE ''%'+isnull(fornecedor, razao)+'%'')'
if @por = 'RC'
begin
    if @documento<>''
        set @sql = @sql + ' and cd_docum_nr_rqco=''''+@documento+''''
        set @sql = @sql + ' and dt_rqco between '''+cast(@dtini as varchar(50))+'''' and
'''+cast(@dtfim as varchar(50))+''''
    end else
        begin
            if @por = 'PC'
                begin
                    if @documento<>''
                        set @sql = @sql + ' and cd_docum_nr_pdco=''''+@documento+''''
                        set @sql = @sql + ' and dt_pdco between '''+cast(@dtini as varchar(50))+'''' and
'''+cast(@dtfim as varchar(50))+''''
                    end else
                        begin
                            if @documento<>''
                                set @sql = @sql + ' and cd_docum_nr_nent=''''+@documento+''''
                                set @sql = @sql + ' and dt_nent_entrada between '''+cast(@dtini as varchar(50))+''''
and '''+cast(@dtfim as varchar(50))+''''
                            end
                        end
                end
            if @semDoc = 'RC'
                set @sql = @sql + ' and cd_docum_nr_rqco is null'
            else if @semDoc = 'PC'
                set @sql = @sql + ' and cd_docum_nr_PDco is null'
            else if @semDoc = 'NE'
                set @sql = @sql + ' and cd_docum_nr_nent is null'
        end
    print (@sql)

exec (@sql)
```

## Stored Procedure RELATORIO\_NOTAS

```
set ANSI_NULLS ON
set QUOTED_IDENTIFIER ON
go

ALTER PROCEDURE [dbo].[relatorio_notas] @estab varchar(4),@texto varchar(50), @nota
varchar(50), @PV varchar(50), @lw varchar(50), @CC varchar(30), @dt_NFini datetime,
@dt_NFfim datetime
AS
BEGIN
    -- SET NOCOUNT ON added to prevent extra result sets from
    -- interfering with SELECT statements.
    SET NOCOUNT ON;

    declare @sql varchar(8000),@sql1 varchar(8000), @sql2 varchar(8000), @sql3
    varchar(8000), @sql4 varchar(8000)

    SET @SQL= 'select nf.cd_estab, nf.cd_docum_nota_sai, nf.cd_docum_nr_nota_sai,
nf.dt_nota_sai, pv.cd_pessoa_comis '
    SET @SQL= @SQL + ' , (select max(isnull(nm_denominacao,nm_fantasia)) from [bra-sql-
rj].sispro_2008v26.dbo.gg_pessoa gg where gg.cd_pessoa=pv.cd_pessoa_comis) as
"solicitante"'
    SET @SQL= @SQL + ' , (SELECT MAX(CD_USUARIO) FROM [bra-sql-
rj].sispro_2008v26.dbo.gc_NSai gg where nf.cd_estab=GG.cd_estab and
nf.cd_docum_nr_nota_sai=GG.cd_docum_nr_nota_sai and
nf.cd_docum_nota_sai=GG.cd_docum_nota_sai) AS USU'
    SET @SQL= @SQL + ' , pv.cd_pessoa_respons (select
max(isnull(nm_denominacao,nm_fantasia)) from [bra-sql-rj].sispro_2008v26.dbo.gg_pessoa
gg where gg.cd_pessoa=pv.cd_pessoa_respons) as "Vendedor"'
    SET @SQL= @SQL + ' , (select max(AN_CGC_CPF) from [bra-sql-
rj].sispro_2008v26.dbo.gg_pessoa gg where gg.cd_pessoa=pv.cd_pessoa_respons) as "CNPJ"'
    SET @SQL= @SQL + ' , nf.cd_pessoa_destinat, (select
max(isnull(nm_denominacao,nm_fantasia)) from [bra-sql-rj].sispro_2008v26.dbo.gg_pessoa
gg where gg.cd_pessoa=nf.cd_pessoa_destinat) as "destinatario"'
    SET @SQL= @SQL + ' , ISNULL((case pv.cd_eoft when ''LW'' then pv.cd_eof else case
pv.cd_eoft_02 when ''LW'' then pv.cd_eof_02 else case pv.cd_eoft_03 when ''LW'' then
pv.cd_eof_03 else case pv.cd_eoft_04 when ''LW'' then pv.cd_eof_04 else case
pv.cd_eoft_05 when ''LW'' then pv.cd_eof_05 else '''' end end end end end),'''')) as LW'
    SET @SQL= @SQL + ' , ISNULL((case pv.cd_eoft when ''CC'' then pv.cd_eof else case
pv.cd_eoft_02 when ''CC'' then pv.cd_eof_02 else case pv.cd_eoft_03 when ''CC'' then
pv.cd_eof_03 else case pv.cd_eoft_04 when ''CC'' then pv.cd_eof_04 else case
pv.cd_eoft_05 when ''CC'' then pv.cd_eof_05 else '''' end end end end end),'''')) as CC'
    SET @SQL= @SQL + ' , ISNULL((case pv.cd_eoft when ''FDC'' then pv.cd_eof else case
pv.cd_eoft_02 when ''FDC'' then pv.cd_eof_02 else case pv.cd_eoft_03 when ''FDC'' then
pv.cd_eof_03 else case pv.cd_eoft_04 when ''FDC'' then pv.cd_eof_04 else case
pv.cd_eoft_05 when ''FDC'' then pv.cd_eof_05 else '''' end end end end end),'''')) as
fdc'
    SET @SQL= @SQL + ' , ISNULL((case pv.cd_eoft when ''CNTR'' then pv.cd_eof else case
pv.cd_eoft_02 when ''CNTR'' then pv.cd_eof_02 else case pv.cd_eoft_03 when ''CNTR'' then
pv.cd_eof_03 else case pv.cd_eoft_04 when ''CNTR'' then pv.cd_eof_04 else case
pv.cd_eoft_05 when ''CNTR'' then pv.cd_eof_05 else '''' end end end end end),'''')) as
CNTR'
    SET @SQL= @SQL + ' , ISNULL((case pv.cd_eoft when ''DEPTO'' then pv.cd_eof else case
pv.cd_eoft_02 when ''DEPTO'' then pv.cd_eof_02 else case pv.cd_eoft_03 when ''DEPTO''
then pv.cd_eof_03 else case pv.cd_eoft_04 when ''DEPTO'' then pv.cd_eof_04 else case
pv.cd_eoft_05 when ''DEPTO'' then pv.cd_eof_05 else '''' end end end end end),'''')) as
DEPTO'
    SET @SQL= @SQL + ' , it.cd_produto, (select ds_produto from [bra-sql-
rj].sispro_2008v26.dbo.gi_produto pr where pr.cd_produto=it.cd_produto) as descricao,
it.an_descr_cpl'
    SET @SQL= @SQL + ' , it.cd_unmed_saida, it.qt_nota_sai_item_prod AS qt_nota_sai_item,
it.vl_unitario, it.vl_total_item, isnull(it.cd_movto_etq_tipo,'''') as tipo'
    SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_nota_sai=it.cd_docum_nr_nota_sai
tr.sq_nota_sai_item=it.sq_nota_sai_item_prod and tr.in_tributo=''IPI''),0) as ipi'
    SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_nota_sai=it.cd_docum_nr_nota_sai
tr.sq_nota_sai_item=it.sq_nota_sai_item_prod and tr.in_tributo=''PIS''),0) as pis'
    SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_nota_sai=it.cd_docum_nr_nota_sai
tr.sq_nota_sai_item=it.sq_nota_sai_item_prod and tr.in_tributo=''ISS''),0) as iss'
    SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
```

```

tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai and
tr.sq_notasai_item=it.sq_notasai_item_prod and tr.in_tributo='IRRF'),0) as irrf'
SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai
tr.sq_notasai_item=it.sq_notasai_item_prod and tr.in_tributo='CSLL'),0) as csll'
SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai
tr.sq_notasai_item=it.sq_notasai_item_prod and tr.in_tributo='ICMS'),0) as icms'
SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai
tr.sq_notasai_item=it.sq_notasai_item_prod and tr.in_tributo='CFINS'),0) as cofins'
SET @SQL= @SQL + ' , isnull((select sum(vl_Trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai
tr.sq_notasai_item=it.sq_notasai_item_prod and tr.in_tributo='INSS'),0) as inss'
SET @SQL1= @SQL1 + ' , tx.tx_texto, case isnull(in_cancel, '') when '' then 'N'
else 'S' end as cancel'
SET @SQL1= @SQL1 + ' , isnull((SELECT max(cd_classif_ipi) FROM [bra-sql-
rj].sispro_2008v26.dbo.gc_notasai_item Impr where impr.cd_estab=it.cd_estab and
imp.cd_docum_nr_notasai=it.cd_docum_nr_notasai
imp.sq_notasai_item Impr=it.sq_notasai_item Prod), '') as ncm'
SET @SQL2= ' from [bra-sql-rj].sispro_2008v26.dbo.gc_notasai_item Prod it, [bra-sql-
rj].sispro_2008v26.dbo.gc_negoc pv, [bra-sql-rj].sispro_2008v26.dbo.gc_notasai nf'
SET @SQL2= @SQL2 + ' left join [bra-sql-rj].sispro_2008v26.dbo.gc_notasai_txt Impr tx
on nf.cd_estab=tx.cd_estab and nf.cd_docum_nr_notasai=tx.cd_docum_nr_notasai'
SET @SQL2= @SQL2 + ' where nf.cd_estab=it.cd_estab and
nf.cd_docum_nr_notasai=it.cd_docum_nr_notasai
nf.cd_docum_notasai=it.cd_docum_notasai and it.cd_negoc=pv.cd_negoc and
it.cd_ESTAB=pv.cd_ESTAB'
SET @SQL2= @SQL2 + ' and isnull(tx.tx_texto, '') like '%'+@texto+'%' '
SET @SQL2= @SQL2 + ' and nf.cd_estab LIKE '%'+@estab+'%' '
SET @SQL2= @SQL2 + ' and (nf.cd_docum_nr_notasai like '%'+@notasai+'%' or
'+@notasai+'%'= '' OR '+@notasai+'%' like '%'+nf.cd_docum_nr_notasai+'%' )'
SET @SQL2= @SQL2 + ' and (it.cd_negoc like '%'+@PV+'%' or '+@PV+'%'= '' or
'+@PV+'%' like '%'+it.cd_negoc+'%' )'
SET @SQL2= @SQL2 + ' and (ISNULL((case pv.cd_eoft when 'LW' then pv.cd_eoft else case
pv.cd_eoft_02 when 'LW' then pv.cd_eoft_02 else case pv.cd_eoft_03 when 'LW' then
pv.cd_eoft_03 else case pv.cd_eoft_04 when 'LW' then pv.cd_eoft_04 else case
pv.cd_eoft_05 when 'LW' then pv.cd_eoft_05 else '' end end end end), '') like
'+@LW+'%' or '+@LW+'%'= '' or '+@LW+'%' like '%'+ISNULL((case pv.cd_eoft when
'LW' then pv.cd_eoft else case pv.cd_eoft_02 when 'LW' then pv.cd_eoft_02 else case
pv.cd_eoft_03 when 'LW' then pv.cd_eoft_03 else case pv.cd_eoft_04 when 'LW' then
pv.cd_eoft_04 else case pv.cd_eoft_05 when 'LW' then pv.cd_eoft_05 else '' end
end end end), ''))+'' )'
SET @SQL2= @SQL2 + ' and (ISNULL((case pv.cd_eoft when 'CC' then pv.cd_eoft else case
pv.cd_eoft_02 when 'CC' then pv.cd_eoft_02 else case pv.cd_eoft_03 when 'CC' then
pv.cd_eoft_03 else case pv.cd_eoft_04 when 'CC' then pv.cd_eoft_04 else case
pv.cd_eoft_05 when 'CC' then pv.cd_eoft_05 else '' end end end end), '') like
'+@CC+'%' or '+@CC+'%'= '' or '+@CC+'%' like '%'+ISNULL((case pv.cd_eoft when
'CC' then pv.cd_eoft else case pv.cd_eoft_02 when 'CC' then pv.cd_eoft_02 else case
pv.cd_eoft_03 when 'CC' then pv.cd_eoft_03 else case pv.cd_eoft_04 when 'CC' then
pv.cd_eoft_04 else case pv.cd_eoft_05 when 'CC' then pv.cd_eoft_05 else '' end
end end end), ''))+'' )'
set @sql2= @sql2 + ' and nf.dt_notasai between '''+cast(@dt_NFinis as varchar(50))+ ''
and '''+cast(@dt_NFfim as varchar(50))+ '' '
SET @SQL3= ' union select nf.cd_estab, nf.cd_docum_notasai, nf.cd_docum_nr_notasai,
nf.dt_notasai, pv.cd_pessoa_comis '
SET @SQL3= @SQL3 + ' , (select max(isnull(nm_denominacao,nm_fantasia)) from [bra-sql-
rj].sispro_2008v26.dbo.gg_pessoa gg where gg.cd_pessoa=pv.cd_pessoa_comis) as
"solicitante" '
SET @SQL3= @SQL3 + ' , (SELECT MAX(CD_USUARIO) FROM [bra-sql-
rj].sispro_2008v26.dbo.gc_NSAI gg where nf.cd_estab=GG.cd_estab and
nf.cd_docum_nr_notasai=GG.cd_docum_nr_notasai
nf.cd_docum_notasai=GG.cd_docum_notasai) AS USU'
SET @SQL3= @SQL3 + ' , pv.cd_pessoa_respons , (select
max(isnull(nm_denominacao,nm_fantasia)) from [bra-sql-rj].sispro_2008v26.dbo.gg_pessoa
gg where gg.cd_pessoa=pv.cd_pessoa_respons) as "Vendedor" '
SET @SQL3= @SQL3 + ' , (select max(AN_CGC_CPF) from [bra-sql-
rj].sispro_2008v26.dbo.gg_pessoa gg where gg.cd_pessoa=pv.cd_pessoa_respons) as "CNPJ" '
SET @SQL3= @SQL3 + ' , nf.cd_pessoa_destinat, (select
max(isnull(nm_denominacao,nm_fantasia)) from [bra-sql-rj].sispro_2008v26.dbo.gg_pessoa
gg where gg.cd_pessoa=nf.cd_pessoa_destinat) as "destinatario" '
SET @SQL3= @SQL3 + ' , ISNULL((case pv.cd_eoft when 'LW' then pv.cd_eoft else case
pv.cd_eoft_02 when 'LW' then pv.cd_eoft_02 else case pv.cd_eoft_03 when 'LW' then

```

```

pv.cd_eof_03 else case pv.cd_eoft_04 when 'LW' then pv.cd_eof_04 else case
pv.cd_eoft_05 when 'LW' then pv.cd_eof_05 else '' end end end end),''') as LW'
SET @SQL3= @SQL3 + ' , ISNULL((case pv.cd_eoft when 'CC' then pv.cd_eof else case
pv.cd_eoft_02 when 'CC' then pv.cd_eof_02 else case pv.cd_eoft_03 when 'CC' then
pv.cd_eof_03 else case pv.cd_eoft_04 when 'CC' then pv.cd_eof_04 else case
pv.cd_eoft_05 when 'CC' then pv.cd_eof_05 else '' end end end end),''') as CC'
SET @SQL3= @SQL3 + ' , ISNULL((case pv.cd_eoft when 'FDC' then pv.cd_eof else case
pv.cd_eoft_02 when 'FDC' then pv.cd_eof_02 else case pv.cd_eoft_03 when 'FDC' then
pv.cd_eof_03 else case pv.cd_eoft_04 when 'FDC' then pv.cd_eof_04 else case
pv.cd_eoft_05 when 'FDC' then pv.cd_eof_05 else '' end end end end),''') as
fdc'
SET @SQL3= @SQL3 + ' , ISNULL((case pv.cd_eoft when 'CNTR' then pv.cd_eof else case
pv.cd_eoft_02 when 'CNTR' then pv.cd_eof_02 else case pv.cd_eoft_03 when 'CNTR' then
pv.cd_eof_03 else case pv.cd_eoft_04 when 'CNTR' then pv.cd_eof_04 else case
pv.cd_eoft_05 when 'CNTR' then pv.cd_eof_05 else '' end end end end),''') as
CNTR'
SET @SQL3= @SQL3 + ' , ISNULL((case pv.cd_eoft when 'DEPTO' then pv.cd_eof else case
pv.cd_eoft_02 when 'DEPTO' then pv.cd_eof_02 else case pv.cd_eoft_03 when 'DEPTO'
then pv.cd_eof_03 else case pv.cd_eoft_04 when 'DEPTO' then pv.cd_eof_04 else case
pv.cd_eoft_05 when 'DEPTO' then pv.cd_eof_05 else '' end end end end),''') as
DEPTO'
SET @SQL3= @SQL3 + ' , nf.cd_condfat, nf.cd_negoc_tipo,
it.cd_trib_oper,it.sq_notasai_servico as sq_notasai_item, it.cd_negoc,
it.sq_negoc_servico'
SET @SQL3= @SQL3 + ' , it.cd_servico, (select ds_produto from [bra-sql-
rj].sispro_2008v26.dbo.gi_produto pr where pr.cd_produto=it.cd_servico) as descricao,
'' as an_descr_cpl'
SET @SQL3= @SQL3 + ' , '' as cd_unmed_saida, it.qt_notasai_servico AS
qt_notasai_item, it.vl_unitario, it.vl_total_item, '' as tipo'
SET @SQL3= @SQL3 + ' , isnull((select sum(vl_trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai and
tr.sq_notasai_item=it.sq_notasai_servico and tr.in_tributo='IPI'),0) as ipi'
SET @SQL3= @SQL3 + ' , isnull((select sum(vl_trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai and
tr.sq_notasai_item=it.sq_notasai_servico and tr.in_tributo='PIS'),0) as pis'
SET @SQL3= @SQL3 + ' , isnull((select sum(vl_trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai and
tr.sq_notasai_item=it.sq_notasai_servico and tr.in_tributo='ISS'),0) as iss'
SET @SQL3= @SQL3 + ' , isnull((select sum(vl_trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai and
tr.sq_notasai_item=it.sq_notasai_servico and tr.in_tributo='ICMS'),0) as icms'
SET @SQL3= @SQL3 + ' , isnull((select sum(vl_trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai and
tr.sq_notasai_item=it.sq_notasai_servico and tr.in_tributo='CFINS'),0) as cofins'
SET @SQL3= @SQL3 + ' , isnull((select sum(vl_trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_trib_item tr where tr.cd_estab=it.cd_estab and
tr.cd_docum_nr_notasai=it.cd_docum_nr_notasai and
tr.sq_notasai_item=it.sq_notasai_servico and tr.in_tributo='INSS'),0) as inss'
SET @SQL4= ' , 0 as margem, 0 as percentual'
SET @SQL4= @SQL4 + ' , tx.tx_texto, case isnull(in_cancel,'') when '' then 'N'
else 'S' end as cancel, '' ncm'
SET @SQL4= @SQL4 + ' , (SELECT MAX(DS_TRIB_OPER) FROM [bra-sql-
rj].sispro_2008v26.dbo.GC_TRIB_OPER OPER WHERE OPER.cd_trib_oper=IT.cd_trib_oper)
DESCCFOF, NF.OB_NSAI_CANCEL'
SET @SQL4= @SQL4 + ' , (select max(an_sit_trib) from [bra-sql-
rj].sispro_2008v26.dbo.gc_nsai_item gcit where it.cd_estab=gcit.cd_estab and
it.cd_docum_notasai=gcit.cd_docum_notasai
and
it.cd_docum_nr_notasai=gcit.cd_docum_nr_notasai
and
it.sq_notasai_servico=gcit.sq_notasai_item) as cst'
SET @SQL4= @SQL4 + ' , NULL, NULL, NULL, NULL'
SET @SQL4= @SQL4 + ' from [bra-sql-rj].sispro_2008v26.dbo.gc_notasai_servico it, [bra-
sql-rj].sispro_2008v26.dbo.gc_negoc pv, [bra-sql-rj].sispro_2008v26.dbo.gc_notasai nf'
SET @SQL4= @SQL4 + ' left join [bra-sql-rj].sispro_2008v26.dbo.gc_notasai_txt Impr tx
on nf.cd_estab=tx.cd_estab and nf.cd_docum_nr_notasai=tx.cd_docum_nr_notasai'
SET @SQL4= @SQL4 + ' where nf.cd_estab=it.cd_estab and
nf.cd_docum_nr_notasai=it.cd_docum_nr_notasai and it.cd_negoc=pv.cd_negoc and
it.cd_ESTAB=pv.cd_ESTAB'
SET @SQL4= @SQL4 + ' and isnull(tx.tx_texto,'') like '''+@texto+''''
SET @SQL4= @SQL4 + ' and nf.cd_estab LIKE '''+@estab+''''
SET @SQL4= @SQL4 + ' and (nf.cd_docum_nr_notasai like '''+@nota+'''' or
'''+@nota+''=''' OR '''+@nota+'' like '''+nf.cd_docum_nr_notasai+''')'

```

```

SET @SQL4= @SQL4 + ' and (it.cd_negoc like '''+@PV+'''' or '''+@PV+''''='''' or
'''+@PV+'''' like '''+it.cd_negoc+''''))'
SET @SQL4= @SQL4 + ' and (ISNULL((case pv.cd_eof when 'LW' then pv.cd_eof else case
pv.cd_eof_02 when 'LW' then pv.cd_eof_02 else case pv.cd_eof_03 when 'LW' then
pv.cd_eof_03 else case pv.cd_eof_04 when 'LW' then pv.cd_eof_04 else case
pv.cd_eof_05 when 'LW' then pv.cd_eof_05 else '''' end end end end end),'''')) like
'''+@LW+'''' or '''+@LW+''''='''' or '''+@LW+'''' like '''+ISNULL((case pv.cd_eof when
'LW' then pv.cd_eof else case pv.cd_eof_02 when 'LW' then pv.cd_eof_02 else case
pv.cd_eof_03 when 'LW' then pv.cd_eof_03 else case pv.cd_eof_04 when 'LW' then
pv.cd_eof_04 else case pv.cd_eof_05 when 'LW' then pv.cd_eof_05 else '''' end end end
end end),''''))+''''))'
SET @SQL4= @SQL4 + ' and (ISNULL((case pv.cd_eof when 'CC' then pv.cd_eof else case
pv.cd_eof_02 when 'CC' then pv.cd_eof_02 else case pv.cd_eof_03 when 'CC' then
pv.cd_eof_03 else case pv.cd_eof_04 when 'CC' then pv.cd_eof_04 else case
pv.cd_eof_05 when 'CC' then pv.cd_eof_05 else '''' end end end end end),'''')) like
'''+@CC+'''' or '''+@CC+''''='''' or '''+@CC+'''' like '''+ISNULL((case pv.cd_eof when
'CC' then pv.cd_eof else case pv.cd_eof_02 when 'CC' then pv.cd_eof_02 else case
pv.cd_eof_03 when 'CC' then pv.cd_eof_03 else case pv.cd_eof_04 when 'CC' then
pv.cd_eof_04 else case pv.cd_eof_05 when 'CC' then pv.cd_eof_05 else '''' end end end
end end),''''))+''''))'
set @sql4= @sql4 + ' and nf.dt_nota_sai between '''+cast(@dt_NFin as varchar(50))+ ''''
and '''+cast(@dt_NFfim as varchar(50))+ ''''

--print (@sql+@sql2)
exec (@sql+@sql1+@sql2+@sql3+@sql4)

-- Insert statements for procedure here
-- SELECT <@Param1, sysname, @p1>, <@Param2, sysname, @p2>
END

```

## Anexo II – Two-Step Cluster results

### Two-Step Cluster – Data Mart Materiais



Valores obtidos utilizando dados reais

Valores obtidos após normalização das variáveis

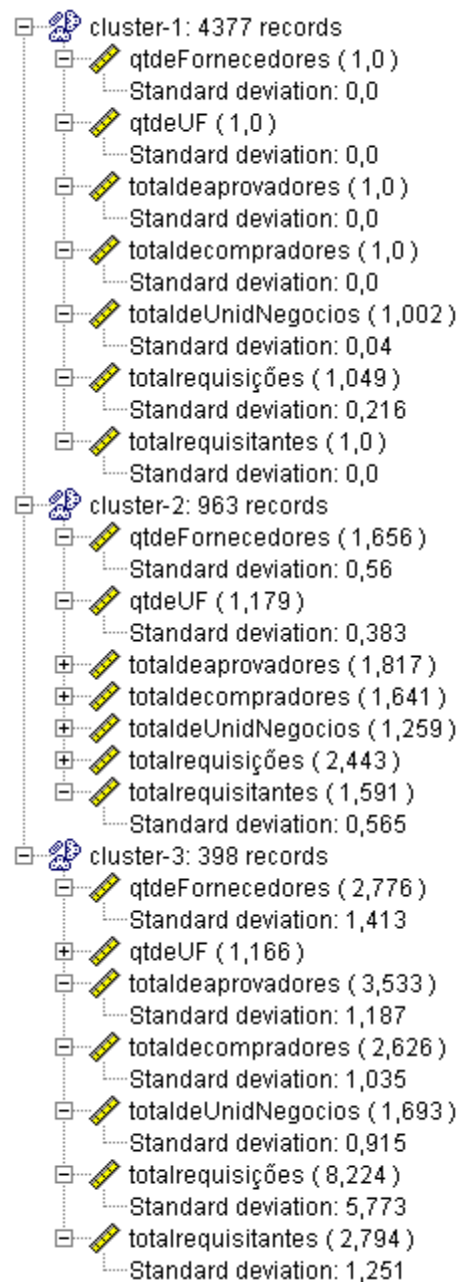
Two-Step cluster – opção selecionada: "Automatically calculate number of cluster"  
(Number of clusters = 2)

Two-Step Cluster – Data Mart Materiais



Number of clusters = 4

Não foi feita normalização dos dados.

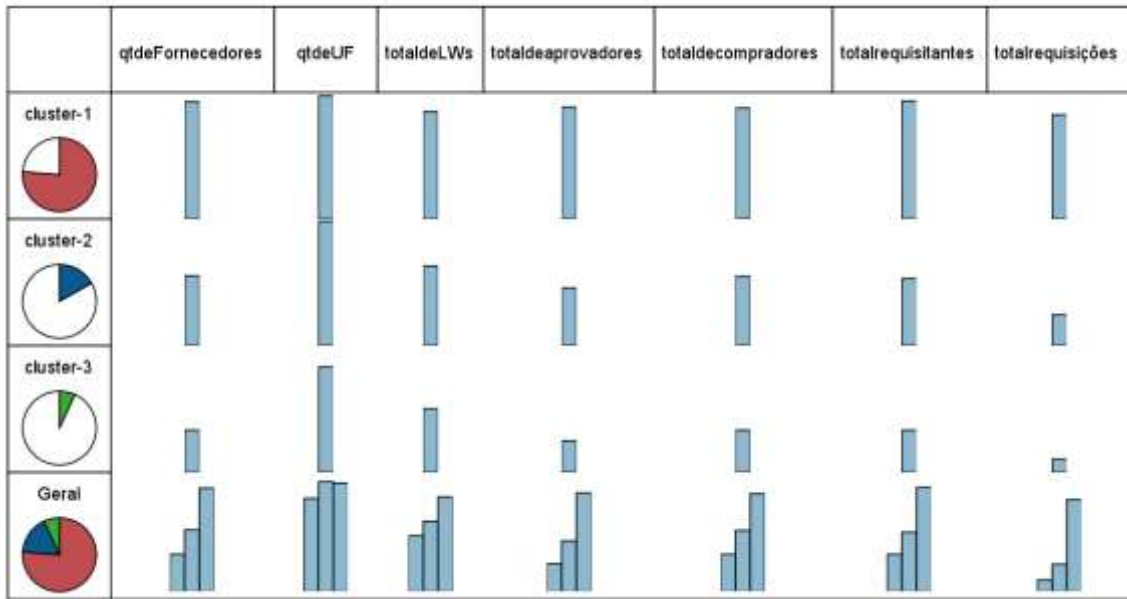


Number of clusters = 3

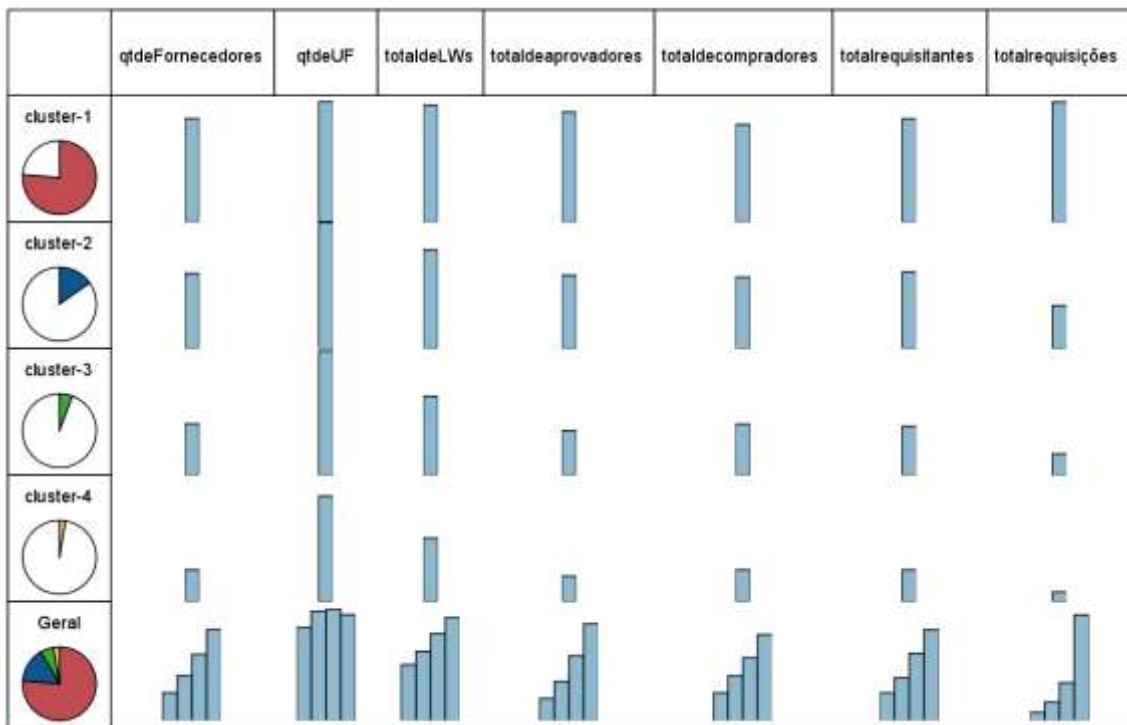
Não foi feita normalização dos dados.

Two-Step cluster – Opção selecionada: "Specify number of clusters"

Two-Step Cluster – Data Mart Materiais

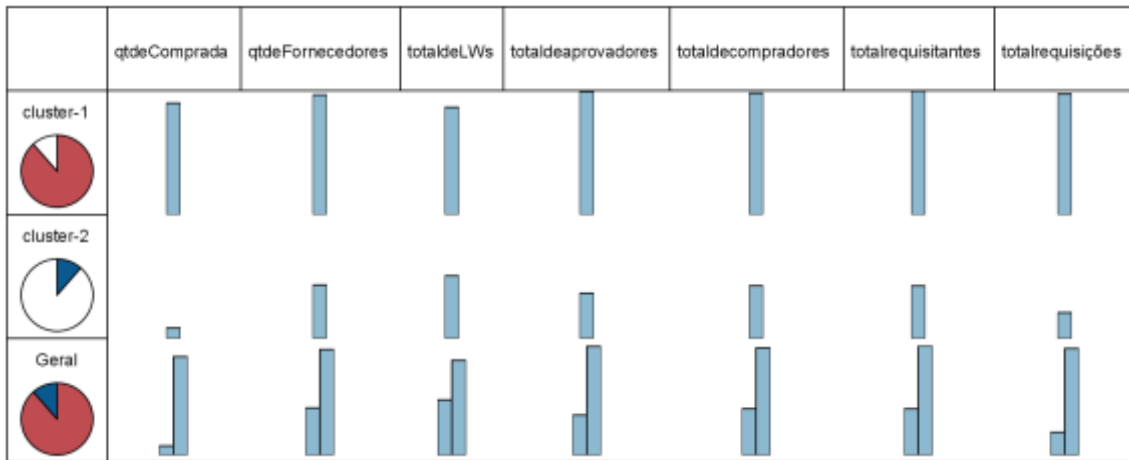


Number of clusters = 3  
 Não foi feita normalização dos dados.

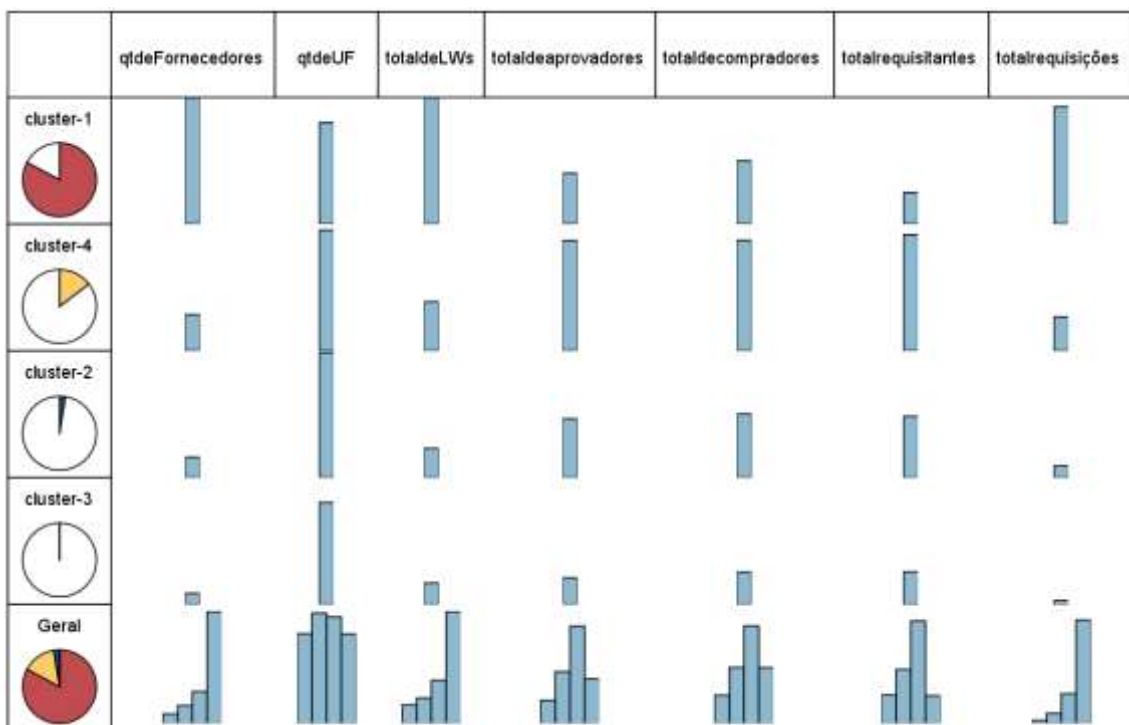


Number of clusters = 4  
 Não foi feita normalização dos dados.

## Anexo III – K-means: Representação gráfica dos clusters (k=2 e k=4)



K-means - Number of clusters = 2



K-means - Number of clusters = 4

## Anexo IV – K-means: Convergência da clusterização

### Number of clusters: 2

Iteration	Error
1	0,838
2	0,33
3	0,191
4	0,102
5	0,041
6	0,027
7	0,062
8	0,027
9	0,01
10	0,007
11	0,001
12	0,0
13	0,0
14	0,0
15	0,0
16	0,0
17	0,0
18	0,0
19	0,0
20	...
30	0,0

### Number of clusters: 3

Iteration	Error
1	0,817
2	0,387
3	0,202
4	0,24
5	0,202
6	0,128
7	0,087
8	0,072
9	0,038
10	0,032
11	0,031
12	0,022
13	0,004
14	0,008
15	0,008
16	0,012
17	0,01
18	0,0
19	0,0
20	...
30	0,0

### Number of clusters: 4

Iteration	Error
1	0,751
2	0,308
3	0,168
4	0,128
5	0,093
6	0,077
7	0,054
8	0,039
9	0,035
10	0,022
11	0,009
12	0,013
13	0,012
14	0,014
15	0,0
16	0,0
17	0,0
18	0,0
19	0,0
20	...
30	0,0

### Number of clusters: 5

Iteration	Error
1	0,751
2	0,306
3	0,17
4	0,128
5	0,093
6	0,077
7	0,054
8	0,039
9	0,032
10	0,022
11	0,009
12	0,008
13	0,01
14	0,009
15	0,008
16	0,003
17	0,0
18	0,0
19	0,0
20	...
30	0,0

## Anexo V – GRI: Data Mart Materiais

	Consequent	Antecedent	Support %	Confidence %
1	codmaterial = 1999001726,000	qtdeComprada < 0,700 and qtdeComprada > 0,325	0,02	100
2	codmaterial = 1399023277,000	VL_TOTAL_PEDIDO > 481400,359 and totalrequisições > 1,500	0,02	100
3	codmaterial = 1399022162,000	qtdeComprada < 0,325	0,02	100
4	codmaterial = 1399021616,000	qtdeComprada < 1,000 and totalrequisições > 3,500	0,02	100
5	codmaterial = 1399014059,000	qtdeComprada > 107445,500	0,02	100
6	codmaterial = 1399013978,000	totaldeUnidNegocios > 5,500	0,02	100
7	codmaterial = 1399012875,000	qtdeUF > 2,500 and totalrequisições < 6,500	0,02	100
8	codmaterial = 1399012333,000	VL_TOTAL_PEDIDO > 481400,359 and totalrequisições < 1,500	0,02	100
9	codmaterial = 1399010845,000	qtdeComprada > 74600,000 and totalrequisições > 15,500	0,02	100
10	codmaterial = 1399009768,000	totalrequisições > 39,500	0,02	100
11	codmaterial = 1399004108,000	totalrequisitantes > 10,000	0,02	100
12	codmaterial = 1399000147,000	totaldeaprovadores > 7,500 and totalrequisições < 10,500	0,02	100
13	codmaterial = 1399000147,000	totalrequisitantes > 8,500 and totalrequisições < 10,500	0,02	100
14	codmaterial = 1199000189,000	qtdeUF > 3,500	0,02	100
15	codmaterial = 1325001027,000	totaldecompradores > 6,500	0,02	100
16	codmaterial = 1999001726,000	qtdeComprada < 0,700	0,03	50
17	codmaterial = 1399023277,000	VL_TOTAL_PEDIDO > 481400,359	0,03	50
18	codmaterial = 1399000147,000	totalrequisitantes > 8,500	0,03	50
19	codmaterial = 1399010845,000	qtdeComprada > 74600,000	0,03	50
20	codmaterial = 1399012333,000	VL_TOTAL_PEDIDO > 481400,359	0,03	50
21	codmaterial = 1399012481,000	VL_TOTAL_PEDIDO < 0,350	0,03	50
22	codmaterial = 1399012498,000	VL_TOTAL_PEDIDO < 0,350	0,03	50
23	codmaterial = 1399012875,000	qtdeUF > 2,500	0,03	50
24	codmaterial = 1399021616,000	qtdeComprada < 1,000	0,05	33,33