FACULDADE DE ECONOMIA E FINANÇAS IBMEC
PROGRAMA DE PÓS-GRADUAÇÃO E PESQUISA EM
ADMINISTRAÇÃO E ECONOMIA

DISSERTAÇÃO DE MESTRADO
PROFISSIONALIZANTE EM ADMINISTRAÇÃO

"DATAMINING USANDO O SODAS: UM ESTUDO DE CASO AONDENAMORO.COM"

EDGARD PEREZ FERNANDES NOGUEIRA

ORIENTADORA: PROFª DRª MARIA AUGUSTA SOARES MACHADO

Rio de Janeiro, 4 de janeiro de 2007

“DATA MINING USANDO O SODAS: UM ESTUDO DE CASO AONDENAMORO.COM”

EDGARD PEREZ FERNANDES NOGUEIRA

Dissertação apresentada ao curso de Mestrado Profissionalizante em Administração como requisito parcial para obtenção do Grau de Mestre em Administração.
Área de Concentração: Administração Geral

ORIENTADORA: PROFª DRª MARIA AUGUSTA SOARES MACHADO

Rio de Janeiro, 4 de janeiro de 2007.

“DATAMINING USANDO O SODAS: UM ESTUDO DE CASO

AONDENAMORO.COM”

EDGARD PEREZ FERNANDES NOGUEIRA

Dissertação apresentada ao curso de Mestrado Profissionalizante em Administração como requisito parcial para obtenção do Grau de Mestre em Administração.
Área de Concentração: Administração Geral

Avaliação:

BANCA EXAMINADORA:

_____

PROFª DRª MARIA AUGUSTA SOARES MACHADO (Orientador)
Instituição: Ibmec

_____

PROF. DR. PAULO SERGIO COELHO
Instituição: Ibmec

_____

PROF. DR. ANILTON SALLES GARCIA
Instituição: Universidade Federal do Espírito Santo

Rio de Janeiro, 4 de janeiro de 2007.

005.741 Nogueira, Edgard Perez Fernandes
N778 Datamining usando o sodas: um estudo de caso Aonde.com / Edgard Perez Fernandes Nogueira. – Rio de Janeiro: Faculdades Ibmec. 2007

Dissertação de Mestrado Profissionalizante apresentada ao Programa de Pós-Graduação em Administração das Faculdades Ibmec, como requisito parcial necessário para a obtenção do título de Mestre em Administração.

Área de Concentração: Administração Geral

1. Mineração de dados 2. Tecnologia da informação

DEDICATÓRIA

Dedico esta dissertação a toda minha família que sempre me apoio muito. Especialmente para minha noiva (Luísa Bertino), pela sua compreensão. Dedico também a toda equipe do Ibmec que sempre me atendeu prontamente, fortalecendo meu conhecimento, principalmente a minha orientadora Maria Augusta Machado, que me auxiliou tanto em minha monografia na graduação quanto nesta dissertação de Mestrado.

RESUMO

Ocorreu nos últimos anos um grande crescimento na capacidade de armazenamento de informações por parte das empresas, o que gerou enormes base de dados. Este fato gerou a oportunidade para o surgimento da técnica de *Data Mining*, que consiste em *softwares* capazes de analisar estes dados e identificar padrões e extrair informações estratégicas para tomada de decisões nas empresas. O SODAS (*Symbolic Official Data Analysis System)* é um *software* livre (gratuito) de *Data Mining*, sendo atualmente distribuído através da internet. Neste estudo é demonstrado um tutorial passo a passo da utilização do referido *software*, assim como um exemplo prático de sua utilização no banco de dados do site de relacionamento AondeNamoro.com, expondo a eficácia do *software* e demonstrando para o meio acadêmico e corporativo uma opção para que seja realizado *Data Mining* a baixo custo.

Palavras Chave: Data Mining, Data WareHouse, Clustering, Objetos Simbólicos, SODAS

ABSTRACT

A great growth in the capacity of storage information of the companies occurred in the last years, that generated enormous database. This fact generated the chance for the sprouting of the technique of Data Mining, that consists of softwares capable to analyze these data and to identify standards and extract strategical information for taking decisions in the companies. SODAS (Symbolic Official Date Analysis System) is a free software of Data Mining, being currently distributed through the internet. In this study a tutorial “step by step” of the use of SODAS will be demonstrated, so as a practical example of its use in the data base of the relationship site AondeNamoro.com, displaying the effectiveness of the software and demonstrating for the corporative and academic an option of a low cost and efficient Data Mining software.

Key Words: Data Mining, Data WareHouse, Clustering, Symbolic Objects, SODAS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ASSO	*Analysis System of Symbolic Official data*
CEREMADE	*Centre De Recherche en Mathématiques de la Décision*
SODAS	*Symbolic Official Data Analysis System*
CRM	*Customer Relationship Management*
KDD	*Knowledge-Discovery in Databases*
FUNDP	*Facultés Universitaires Notre-Dame de la Paix*
SGBD	Sistema de Gerenciamento de Banco de Dados
ODBC	*Open Data Base Connectivity*
SDA	*Symbolic Data Analysis*
SQL	*Structured Query Language*

SUMÁRIO

1 INTRODUÇÃO

O AondeNamoro.com é um *site* brasileiro de relacionamento, que utiliza a tecnologia para aproximar pessoas, com as mais diversas necessidades, como por exemplo: amizade, namoro, relacionamento casual e relacionamento sério.

O SODAS (*Symbolic Official Data Analysis System*) é um *software* livre Francês, o qual possui diversas funcionalidades que podem auxiliar na utilização de *Data Mining* por empresas. A intenção ao utilizar o SODAS é demonstrar como, de forma simples e a baixo custo, muitas empresas podem utilizar o *Data Mining* para conhecer melhor seu negócio e seus clientes.

O SODAS é um *software* existente há pouco tempo e que não é comercializado, sendo distribuído gratuitamente através da internet. O programa ainda não é amplamente utilizado e possui uma divulgação ainda restrita ao meio acadêmico especializado em Estatística ou *Data Mining*.

A empresa escolhida para o estudo de caso o site relacionamento AondeNamoro.com possui uma grande diversidade de dados, o que facilita a utilização de técnicas de *Data Mining* para identificar padrões e tendências, o que certamente enriquece este trabalho.

O AondeNamoro.com está em um processo de grandes mudanças estratégicas na empresa, e neste momento precisando extrair conhecimento das informações que já possui de seus usuários. As mudanças que serão realizadas ocorrerão tanto no *layout* do site, forma de divulgação (propaganda) e na forma de receita atual da empresa. Estas mudanças possuem como objetivo aumentar a competitividade da empresa, ganhar uma maior fatia do mercado e ampliar receita.

Através da utilização do *software* SODAS, é realizado o *Data Mining* na base de dados oferecida pelo site AondeNamoro.com, sendo desenvolvido um modelo para ampliar a competitividade da empresa e lhe oferecer informações estratégicas sobre seu banco de dados. Uma contribuição para o meio acadêmico é a demonstração exemplificada de uma ferramenta gratuita e de excelente desempenho.

Estrutura do Trabalho

A dissertação encontra-se estruturada em 5 capítulos, além deste e da lista de referências bibliográficas:

- Capítulo 2 Revisão de literatura dos principais conceitos acadêmicos utilizados nesta dissertação, tais como: *Data Mining*, *Data Warehouse* e *Clustering.*
- Capítulo 3 Contempla informações relevantes sobre o SODAS, com sua origem, foco e Análise de Dados Simbólicos (SDA).
- Capítulo 4 Mineração de dados da base de dados do AondeNamoro.com, buscando identificar informações relevantes para a empresa em questão.
- Capítulo 5 Conclusão do trabalho, vantagens e desvantagens do *software* SODAS, recomendações a empresa, estudos futuros e limitações do trabalho.

1.1 METODOLOGIA DA PESQUISA

1.1.1 JUSTIFICATIVA E CONTEXTUALIZAÇÃO DO PROBLEMA DE PESQUISA

Atualmente existem diversas ferramentas de *Data Mining*, entretanto poucas destas são gratuitas. Através deste estudo, é utilizada a ferramenta SODAS para efetuar o *Data Mining* no banco de dados da empresa AondeNamoro.com.

O AondeNamoro.com é um site de relacionamento na Internet, oferecido pelo mecanismo de busca Aonde.com. O intuito principal do site é criar um meio seguro, interativo e divertido de aproximar pessoas, para quaisquer que sejam suas intenções (amizade, namoro, casamento, etc.).

1.1.1.1 Por que a escolha do AondeNamoro.com

Pelas suas característica, é esperado que o AondeNamoro.com trata-se de um excelente estudo de caso interessante, devido ao mesmo tratar-se de um serviço de relacionamento totalmente via *Web* o que traz diversos benefícios para o estudo, como por exemplo:

- Banco de Dados On-line – Por tratar-se de um serviço completamente via *Web*, todas as informações geradas no site são armazenadas em banco de dados, facilitando assim o acesso às informações;
- Volume de dados – Devido ao AondeNamoro.com possuir uma quantidade de registros considerável, tendo em vista que são mais 30.000 pessoas cadastradas no serviço;

- Quantidade de Variáveis disponíveis – Sendo um site de relacionamento o AondeNamoro.com possui uma enorme gama de informações fornecidas por seus usuários no momento de seu cadastro. (Tabelas 2, 3 e 4);
- Integração com o SODAS – A empresa trabalha com banco de dados MYSQL, que é um banco de fácil utilização e integrado com o SODAS;
- Interesse da empresa com o projeto – O AondeNamoro.com deseja efetuar grandes mudanças estratégicas, e para estas necessita conhecer melhor seus dados, oferecendo assim grande apoio no sucesso da pesquisa.

O AondeNamoro.com planeja efetuar grandes mudanças em sua estrutura nestes próximos meses, mudanças tanto em seu layout, fontes de receita e na forma de divulgação utilizada atualmente. Muitas destas mudanças na estrutura do site e empresa já tiveram início, principalmente quanto ao layout e tecnologias, e provavelmente até o final do estudo outras mudanças já terão sido concretizadas.

Este estudo busca auxiliar o AondeNamoro.com nestas grandes mudanças estratégicas, quanto aos serviços oferecidos no site, e também a identificar as melhores oportunidades para efetuar divulgação da empresa.

A principal mudança estratégica a ser realizada é no modelo de receita utilizado pela empresa e é principalmente no auxilio a esta mudança que este estudo está concentrado.

Atualmente os grandes *sites* de relacionamento na internet brasileira possuem o mesmo modelo de receita, onde o site é gratuito, entretanto para o usuário utilizar grande parte das funcionalidades, deve efetuar o pagamento de uma assinatura (mensal, trimestral ou semestral).

O usuário, após cadastrar-se no site de relacionamento, recebe como bônus alguns dias gratuitos para utilização de todas as funcionalidades, prazo este que varia de 7 a 15 dias. Segundo dados da empresa, normalmente quando expirado este prazo, grande parte dos usuários interrompem a utilização do serviço, pois não efetuam a assinatura. O usuário que não efetua o pagamento da assinatura não consegue ler as mensagens que recebe de outros usuários, o que normalmente é desmotivante, pois não permite mais a comunicação entre os mesmos.

O mercado de *sites* com serviço de relacionamento na internet brasileira nos últimos anos teve um grande aumento, e possui um grande *player* no mercado, chamdo ParPerfeito.com.br, que possui uma quantidade de usuários cadastrados, sendo esta bem superior aos demais concorrentes.

Atualmente, grande parte dos *sites* de relacionamento possuem as mesmas funcionalidades para os usuários (envio de mensagens, elogios, chat, etc.), tornando o mercado de difícil diferenciação. O AondeNamoro.com deseja se diferenciar de seus concorrentes mudando seu modelo de receita.

O AondeNamoro.com, pretende adotar o mesmo modelo de receita efetuado por alguns *sites* internacionais, onde o usuário não paga uma assinatura para utilizar o serviço, pois a receita do site é oriunda exclusivamente de publicidade exposta no site.

Efetuando esta mudança o AondeNamoro.com espera que seus usuários que interromperam a utilização do site devido a necessidade de pagar a assinatura voltem a utilizar o mesmo e também conquistar novos clientes, uma vez que será mais atrativo perante os concorrentes.

Para que esta mudança seja realizada com sucesso e da forma mais lucrativa possível o AondeNamoro.com necessita identificar quais são os grupos (*clusters*) de usuários existentes em seu serviço. O AondeNamoro.com poderá identificar os usuários que acessam mais freqüentemente o serviço e conseqüentemente mais lucrativos, pois são os que visualizam mais publicidades dentro do site e o grupo de usuários menos lucrativos para empresa, ou seja, os que acessam pouco o site. Sendo assim torna-se necessário identificar os usuários que fornecem mais impressões[1] para o *site*.

O AondeNamoro.com também efetuará uma grande mudança em sua forma de divulgação, onde atualmente é principalmente efetuada através do mecanismo de busca Aonde.com, que encontra-se entre os 100 maiores *sites* de língua portuguesa (TOP, 2006). Esta divulgação auxiliou bastante a entrada do AondeNamoro.com no mercado, porém o planejamento é que após a análise realizada nesta dissertação, a empresa possua informações para investir de melhor forma em divulgação através de outros meios e parceiros.

Com a identificação de grupos o AondeNamoro.com poderá mudar a sua forma de divulgação, pois terá como identificar parceiros específicos e criar mensagens específicas para atingir cada tipo de *clusters* de usuários do serviço, fazendo assim com que a empresa utilize mais racionalmente sua verba de publicidade.

Demonstrando de forma resumida os objetivos do AondeNamoro.com com o estudo são:

- Identificação os diversos tipos de *Clusters* de usuários;
- Conhecer os segmentos de usuários do site e entender suas necessidades;
- Identificar os *clusters* de usuários mais lucrativos;

[1] Impressões é o termo utilizado pelo AondeNamoro.com para definir a quantidade de páginas navegadas no site por um usuário. Quanto maior a quantidade de impressões maior a possibilidade de o usuário gerar receita.

- Identificar os *clusters* de usuários menos lucrativos;
- Apoio na decisão sobre quais regiões (estados) devem obter maior fatia da verba de publicidade;
- Ampliar a competitividade do site, oferecendo soluções personalizadas para seus usuários;
- Incentivar o contato de usuários de *clusters* semelhantes, que teoricamente possuem uma maior possibilidade de gerar um relacionamento.

1.1.2 OBJETIVOS DA PESQUISA

Objetivos Gerais

- Efetuar um estudo acadêmico do *Software* livre Francês SODAS, utilizando o mesmo em um caso prático, objetivando demonstrar que é viável a utilização de ferramentas livres para *Data Mining* em base de dados empresariais.
- Ampliar o campo de *Data Mining* para pequenas corporações que nem sempre possuem condições de arcar com o custo de licença dos atuais *software*s do mercado.
- Orientar a importância da utilização do *Data Mining* para competitividade das empresas.

Objetivos Específicos

A seguir podem ser observados alguns objetivos específicos, que se deseja atingir com este estudo.

- Criar um modelo de *Data Mining* para a empresa estudada AondeNamoro.com.

- Localizar no Banco de dados do AondeNamoro.com informações que possam ser relevantes para empresa, oferecendo assim uma retribuição à base concedida para o estudo.
- Criação de um tutorial que auxiliará estudos futuros sobre *Data Mining* com o *software* SODAS.

1.1.3 METODOLOGIA E RELEVÂNCIA DA PESQUISA

A metodologia utilizada nesta dissertação é através do exemplo prático da utilização do SODAS para *Data Mining* no banco de dados da empresa estudada AondeNamoro.com.

Através do exemplo prático, pretende-se demonstrar a eficácia do SODAS e demonstrar como o *Data Mining* com o referido *software* pode acrescentar para o meio corporativo. São utilizados os principais métodos do SODAS relacionados a Estatística Descritiva e Clustering, sendo demonstrados quais tipos de resultados podem ser alcançados com estes métodos.

Durante o estudo é demonstrado passo a passo a utilização do SODAS, sendo criado um tutorial, exposto nos apêndices A e B deste trabalho, da utilização do *software* por pesquisadores ou profissionais que desejarem conhecer melhor o aplicativo.

Este estudo possui como relevância para o meio acadêmico e empresarial, demonstrar como é viável implementar um *Data Mining* a baixo custo, e assim ampliar as técnicas de *Data Mining* também para pequenas corporações e em maior escala para o meio acadêmico.

1.2 LIMITAÇÕES DO ESTUDO

- Utiliza somente alguns métodos do SODAS para efetuar o *Data mining* da empresa estudada*;*
- O tutorial passo a passo não demonstra todos os módulos do SODAS, somente os necessários para o estudo realizado;
- As informações e *clusters* gerados não podem ser extrapolados para todos os *sites* da internet ou todos os *sites* de relacionamento, tornado-se necessário a análise caso a caso.

2 REVISÃO DE LITERATURA

Durante a Revisão de Literatura, a principal preocupação é relacionada com as definições acadêmicas dos principais assuntos abordados durante esta dissertação, sendo estes: *Data Warehouse*, *Data Mining* e *Clustering*.

2.1 *DATA WAREHOUSE*

Sempre que se trabalha com *Data Mining*, não se pode deixar de ressaltar a importância do *Data Warehouse*. Segundo Weiss (1998) o surgimento do *Data Warehouse* pode ser considerado o fato que iniciou a revolução no tratamento de grandes massas de dados, pois o mesmo possibilitou a centralização dos dados de apoio a decisão das empresas.

Segundo Thuraisingham (1998) o *Data Warehouse* é uma das tecnologias de administração de dados que dá suporte ao *Data Mining*. O *Data Warehouse* é uma ferramenta que possibilita aos usuários acessarem facilmente informações que estariam distribuídas em diversas base de dados, mas que estão de forma integrada no *Data Warehouse*.

De acordo com Berson et all (1999) *Data Warehouse* é a utilização de tecnologias visando a integração eficaz de bases de dados em um ambiente que permite o uso estratégico destes dados. Estas tecnologias incluem: sistemas relacionais de gerenciamento de banco de dados, arquitetura cliente/servidor, repositórios, interfaces gráficas, entre outros.

No ponto de vista de Han e Kamber (2001) *Data Warehouse* é um repositório de informações coletadas de diversas fontes diferentes, armazenadas de forma unificada e normalmente no mesmo local. Os *Data Warehouse* são construídos através de limpeza, transformação e integração de dados, onde as informações são atualizadas e renovadas de tempos em tempos

(por exemplo semanalmente). As informações armazenadas nos mesmos possuem como foco facilitar o processo de decisão, sendo normalmente divididas por assuntos (consumidores, fornecedores, entre outros) e preparadas para trabalhar com informações "históricas". O processo de criação e utilização de um *Data Warehouse* pode ser observado na figura 1.

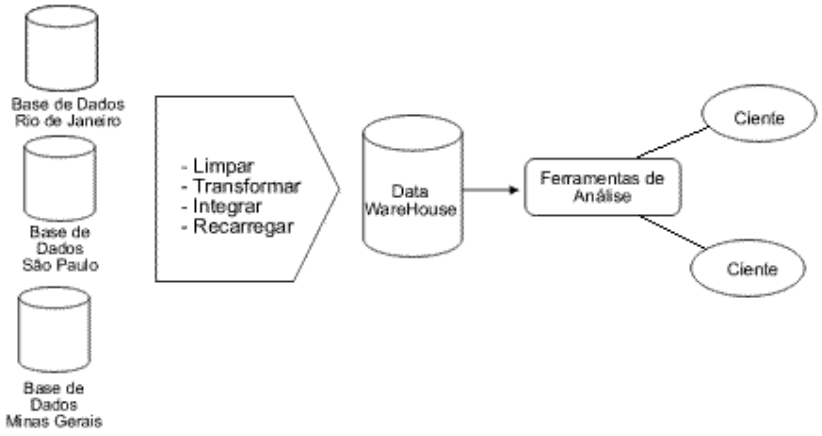


Figura 1 – Processo de criação e utilização de *Data Warehouse* (HAN; KAMBER, 2001, Pág 13)

A técnica de *Data Warehouse* é de suma importância para o desenvolvimento de um *Data Mining* e para empresa poder extrair informações estratégicas de seus dados. Ao desenvolver um *Data Warehouse*, deve ser analisada a qualidade dos dados, tendo em vista que os mesmos são orientados por assuntos e com viés estratégico. A qualidade dos dados deve ser mantida durante o processo de limpeza, transformação, integração e atualização dos dados, visando o banco de dados ter uma alta confiabilidade e integridade de suas informações e desempenho.

Durante o processo de *Data Warehouse* o responsável pelo mesmo deve estar atento a usabilidade, fazendo com que ele seja "prático e fácil" de ser trabalhado, possua escabilidade, tendo em vista que *Data Warehouse* devem ser preparados para uma crescente quantidade de

informações ao longo do tempo, e confiabilidade, tendo em vista que as informações contidas no mesmo devem ser livres de erros e falhas para diminuir os riscos de decisões estratégicas.

Outras informações sobre *Data Warehouse* podem ser localizadas em: (HAN; KAMBER, 2001) e (BERSON et all, 1999)

2.2 *DATA MINING*

Com o surgimento do *Data Warehouse* armazenando grandes massas de dados de forma centralizada e com o crescimento na capacidade de processamento dos computadores, surgiu uma oportunidade para desenvolvimento de tecnologia e ferramentas para extrair informação destes dados. Surgiu assim o *Data Mining*, que foi implementado com muito sucesso, principalmente em "*database marketing"* e sistema anti-fraude. (KENNEDY et all, 1998)

Segundo Silva (2006) no Brasil, tanto no meio acadêmico como no empresarial, as técnicas de mineração de dados ainda não são muito utilizadas. Normalmente, nos enormes bancos de dados existentes são realizadas somente consultas simples, fazendo com que todo o potencial de conhecimento que pode ser extraído destes dados fique ignorado.

Tendo em vista a possibilidade do *Data Mining* obter ótimos resultados com "*database marketing*", pode-se identificar algumas oportunidades de utilizar o mesmo no estudo de caso, buscando alavancar para o AondeNamoro.com alguns pontos de marketing ressaltados por Kennedy et all (1998), como por exemplo:

- *Response Modeling* – Identificar e prever comportamento dos consumidores baseado em dado histórico, demográfico, geográfico e estilo de vida;

- *Cross-selling* – Ampliar a venda de outros produtos ou serviços para a base de consumidores existentes;
- *Customer Valuation* – Prever a receita que pode ser gerada pelo consumidor dentro de um determinado tempo, baseado em informações históricas;
- *Segmentation and profiling* – Ampliar o conhecimento dos clientes, através da análise de dados, buscando compreender melhor os segmentos existentes.

De acordo com Berson et all (1999) *Data Mining* em sua definição mais simples é uma forma automática de detectar padrões relevantes em um banco de dados. Por exemplo, o *Data Mining* sendo utilizado no CRM (*Customer Relationship Management*) para prever atitudes de consumidores. Berson et all (1999) em uma definição mais formal e ampla, definem o *Data Mining* como o processo de descoberta de novas correlações, padrões e tendências significantes, através de grandes base de dados armazenadas em *Data Warehouse*.

Segundo Silva (2006), Mineração de dados é uma etapa do emergente processo de Descoberta de Conhecimento no Banco de Dados, conhecido como KDD (*Knowledge-Discovery in Databases*). A mineração de dados utiliza métodos para localizar padrões nos dados, utilizando parâmetros específicos do algoritmo para cada tarefa desejada. Para que a mineração de dados obtenha sucesso é necessário que os dados possuam uma boa qualidade e que tenham sido tratados anteriormente (limpos, sem inconsistências, entre outros).

De acordo com Han e Kamber (2001) o *Data Mining* surgiu de uma crescente necessidade ocorrida nas corporações, onde com a ampliação da tecnologia, ocorreu uma “explosão” na quantidade de dados armazenados, gerando assim uma grande riqueza de dados nas organizações, que ficaram “ricas” em dados e “pobres” em informação e o *Data Mining* surgiu para tentar reduzir este *gap*.

Han e Kamber (2001) definem em sua forma mais simples que *Data Mining* significa extrair conhecimento de grandes bases de dados, acreditando inclusive que o nome mais apropriado para *Data Mining* (Mineração de Dados) deveria ser "*knowledge mining from data*" (Mineração de conhecimento dos dados). Pode-se verificar através da figura 2 uma arquitetura típica de *Data Mining.*

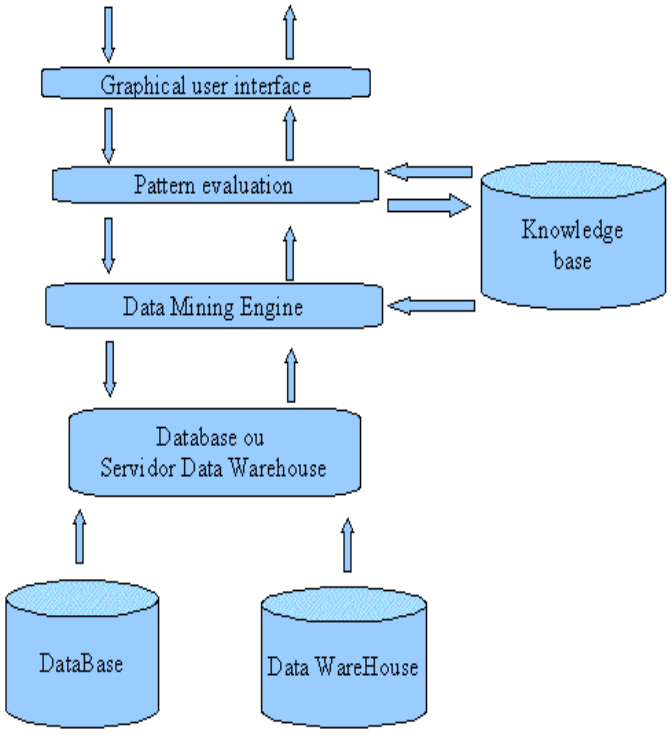


Figura 2 – Arquitetura de um *Data Mining* (HAN; KAMBER, 2001, Pág 8)

Rud (2001) expõe uma definição mais comercial do *Data Mining* onde ele é definido como um termo que engloba uma grande gama de técnicas usadas por várias indústrias, que amplia sua competição e gera mais resultados para as empresas, sendo o mesmo essencial para a manutenção da competitividade das empresas em todo ciclo de contato com os consumidores.

Segundo Thuraisigham (1998), o *Data Mining* é um novo método, composto de ferramentas e técnicas, para resolver problemas relacionados a dados. Estes problemas já existiam a muitos

anos, entretanto agora com o *Data Mining* são disponibilizados técnicas e métodos mais apurados e capazes de analisar uma quantidade enorme de dados.

Outras informações sobre *Data Mining* podem ser localizadas em (BERSON et all,1999), (HAN; KAMBER, 2001)

2.3 *CLUSTERING*

Clustering é uma técnica de estatística multivariada que possui como objetivo dividir "N" elementos com "p" atributos em "k" grupos (*cluters).* Sendo os "k" grupos com elementos homogêneos e os "k" grupos devem ser heterogêneos entre si.

Segundo Press (1972), estatística multivariada tem como foco estudar variáveis que possuam correlação com outra(s) variável(is). Partindo do pressuposto que se existe correlação entre duas variáveis, todo conhecimento adquirido sobre uma pode gerar algum conhecimento sobre a outra variável.

Segundo Hair et all (1995) *Clustering* é uma técnica analítica para desenvolvimento de subgrupos de elementos ou assuntos, tendo como objetivo específico classificar uma amostra em grupos exclusivos de elementos com grande similaridade.

Press (1972) define *clustering* como sendo diversas técnicas para agrupar elementos multidimensionais de acordo com uma variedade de critérios de homogeneidade e heterogeneidade entre os elementos.

Ribeiro (2005) identifica diversas funções para o *Clustering*, como, por exemplo: identificação de segmentação de base de usuários. No setor de seguros pode ser utilizada para análise de risco, entre outros.

Segundo Kaufman e Rousseeuw (1990) *clusters* podem ser classificados em dois métodos, sendo estes:

- Não hierárquico (particionamento) - Neste método são construídos k *clusters*, onde os elementos são divididos em k grupos, e estes elementos possuem todas as características para estarem nestes grupos. Neste tipo de classificação cada cluster possui ao menos um elemento e cada elemento pertence à somente um grupo. O valor de "k" é estipulado pelo usuário que está desenvolvendo o modelo. Sendo assim, é aconselhável que sejam efetuados testes com diferentes valores para "k".

- Hierárquico - Este método possui duas abordagens, sendo a primeira chamada de Aglomerativo (*Botton-up*) e a segunda de divisivo (*Top-Down*).

A abordagem de Aglomerativa (*Botton-up*) inicialmente distribui cada elemento em um *cluster*, fazendo com que a quantidade de *cluster* "k" seja a mesma quantidade de elementos. Ocorre um processo de agrupamento dos clusters até que todos os elementos pertençam a somente um *cluster*. (METZ, 2005)

A abordagem divisiva (*Top-Down*) inicialmente todos os elementos encontram-se agrupados dentro de somente um *cluster* (k = 1) e ocorrem divisões até que seja localizado algum parâmetro ou critério para a parada do algoritmo.

As diferenças das duas abordagens podem ser visualizadas graficamente através da figura 3.

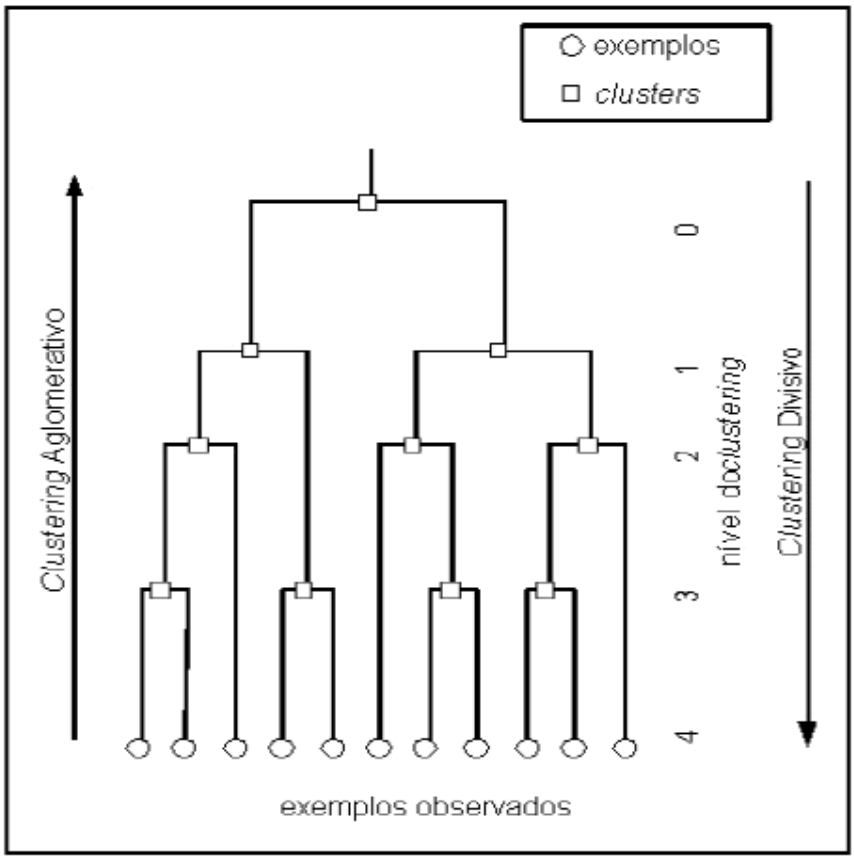


Figura 3 – *Clusters* Divisivo e Aglomerativo (METZ, 2005, Pág 2)

No estudo de caso apresentado, é utilizado principalmente o método de *Clustering* do SODAS, para efetuar a mineração dos dados do AondeNamoro.com, buscando encontrar os *clusters* dos mais diversos segmentos de usuários, fornecendo informações valiosas para a empresa expandir sua competitividade.

3 SODAS

Neste capítulo são abordadas informações relevantes sobre o SODAS (*Symbolic Official Data Analysis System*), para que se possa compreender melhor sua origem, o que são objetos simbólicos e os módulos disponíveis atualmente no *software*.

3.1 ORIGEM DO SODAS

O SODAS é um *software* livre Francês de *Data Mining*, desenvolvido pelo departamento CEREMADE (*Centre De Recherche en Mathématiques de la Décision*) da Universidade de Dauphine na França, desenvolvido para *European Esprit (*Projeto número 20821). O SODAS inicialmente enfrentou dificuldades em sua divulgação, tendo em vista que suas primeiras versões e tutoriais foram desenvolvidos em Francês.

Com o lançamento da versão 1.2 do SODAS (desenvolvida em Inglês) e inclusive com tutoriais e exemplos neste idioma, ocorreu uma ampliação na divulgação da ferramenta. Atualmente o SODAS está em sua versão 2.5, onde ocorreu uma maior profissionalização do produto, principalmente em um dos quesitos mais importantes que é a possibilidade do *Software* importar dados de diferentes tipos de banco de dados, onde nas outras versões também era possível, entretanto de forma mais complexa.

O SODAS por tratar-se de um *software* sem fins lucrativos e sem muitos recursos financeiros, a divulgação do mesmo é realizada praticamente somente no meio acadêmico e através de *Workshops*. (DISSEMINATION, 2006)

De acordo com Objectives (2006) atualmente o projeto do SODAS é mantido pela *ASSO Project (Analysis System of Symbolic Official Data)*, do setor de Informática da FUNDP (*Facultés Universitaires Notre-Dame de la Paix*) na Bélgica, sendo a Sra. Anne de Baenst gerente do *Software*. Maiores informações sobre os responsáveis do desenvolvimento inicial do SODAS podem ser localizadas em (PARTICIPANTS, 2006) e informações sobre os atuais responsáveis, podem ser localizadas em (WORKPACKAGES, 2006; CONSORTIUM, 2006)

Tendo em vista que o SODAS é um *software* desenvolvido no meio acadêmico, existe uma grande preocupação em seus desenvolvedores em ter um respaldo cientifico sobre a tecnologia utilizada no mesmo. Pode-se verificar em (SCIENTIFC, 2006) maiores informações sobre os conceitos acadêmicos utilizados no mesmo. Nesta dissertação busca-se contribuir com a equipe do SODAS com um exemplo prático real da utilização do *software*, e também um tutorial passo a passo da utilização do mesmo, disponível nos apêndices A e B.

Atualmente o SODAS possui as seguintes características e metas para o ASSO *Project*: (FEATURES, 2006)

- Cria, prepara e modela conceitos estatísticos por Objeto Simbólico;
- Desenvolve dados simbólicos de uma base de dados relacional;
- Auxilia a análise de dados estratégicos e administrativos;
- Assegura a qualidade de resultados estatísticos, para usuários que necessitam de informações confiáveis;
- Fácil interação homem/máquina (*user-friendly*);

- Criar novos métodos que auxiliem no desenvolvimento dos modelos e ajudem na análise estatística;
- Adicionar novos métodos de classificação supervisionada ou não supervisionada, que tenham como *input* dados simbólicos e como *output* objetos simbólicos.
- Criação de ferramentas para analisar a qualidade, estabilidade e robustez dos métodos.

3.2 ANÁLISE DE DADOS SIMBÓLICOS

Uma das principais características do SODAS é possuir a grande habilidade em trabalhar com Análise de Dados Simbólicos (*Symbolic Data Analysis* – SDA).

Segundo Carvalho (2006) SDA, está relacionada com análise multivariada, reconhecimento de padrões e inteligência artificial, desenvolvendo métodos para dados descritos por variáveis multivaloradas, sendo a mesma uma nova abordagem na área de descoberta de conhecimento.

De acordo com Agentes (2006) a SDA realiza a extração de conhecimentos de grandes bases de dados, sendo este conhecimento modelado através de dados mais complexos, chamados de "dados simbólicos".

Para Diday (2006) dados simbólicos são gerados quando grandes quantidades de dados são organizados. A necessidade desta organização pode surgir por diversos motivos, como, por exemplo, uma "busca" em um banco de dados que retorne categorias e variáveis.

De acordo com Bezerra (2006) os dados simbólicos são mais complexos que dados usuais, pois apresentam variação interna e são estruturados.

A SDA possui ferramentas capazes de trabalhar com dados complexos, agregados, relacionais e de alto nível, onde as entradas de tabela de dados são conjuntos de categorias ou de números, intervalos ou distribuições de probabilidade associadas a regras e taxonomias. (BEZERRA, 2006)

De acordo com Diday (2006) tabelas de objetos simbólicos, são o principal *input* para Análise de Dados Simbólicos. Estas tabelas são desenvolvidas da seguinte forma: as colunas da tabela são as "variáveis simbólicas" utilizadas para descrever os indivíduos, e as linhas são as "descrições simbólicas" destes indivíduos, sendo esta variável quantitativa ou categórica.

Os objetos simbólicos possuem a capacidade de armazenar uma maior variedade de informações e conceitos, sendo assim melhores para representarem "conceitos" do que indivíduos. (MENESES, 2006)

> Os objetos simbólicos são mais bem adaptados do que os objetos usuais para descrever grupos de indivíduos, levando em conta a variabilidade enquanto disjunção de valores relativos a uma variável. Os objetos simbólicos são representados geometricamente como hiper-cubos onde cada dimensão corresponde aos valores assumidos por cada variável. A presença de regras de dependências entre as variáveis pode restringir e / ou reduzir as dimensões do espaço de descrição dos objetos simbólicos. (CARVALHO, 2001, pág.4)

De acordo com Bezerra e Queiroz (2006) o sodas cria suas tabelas de dados simbólicos a partir de base de dados relacionais.

O conceito de Análise de Objetos Simbólicos utilizado pelo SODAS, é muito interessante para a empresa estudada, tendo em vista que seu banco de dados é totalmente relacional, o que facilita a criação de tabelas de objetos simbólicos. Sendo assim, utilizando o SODAS a empresa estudada pode utilizar uma ferramenta que possui a capacidade de interagir com uma grande quantidade de informações e também a capacidade de interagir com um banco relacional.

Maiores informações sobre Análise de Dados Simbólicos podem ser obtidas em: (DIDAY, 2006) e (CARVALHO, 2006)

3.3 MÓDULOS DO SODAS

As funcionalidades do SODAS são separadas por módulos, o que apresenta diversas vantagens. Primeiramente, faz com que o *software* seja mais *user-friendly,* uma vez que se torna mais fácil localizar qual funcionalidade se deseja utilizar. Outro aspecto positivo desta abordagem é que devido ao mesmo ser um *software* livre, isto gera a possibilidade de outros desenvolvedores implementarem módulos adicionais, gerando assim uma grande possibilidade de crescimento para o *software* no médio prazo.

Na versão 2.5 do SODAS, que foi utilizada para esta dissertação, podem ser encontrados os seguintes métodos ou módulos (tabela 1): (SODAS, 2004).

Módulo	Objetivo
DSTAT	Estatística Descritiva
VIEW	Visualizador de Objeto Simbólico
DISS	Medidas de não semelhança
MATCH	Operadores de cruzamento
DIV	Classificação por divisão
HIPYR	Cluster Hierárquico e Piramidal
SCLUST	Cluster Dinâmico
DCLUST	Algoritmo de Cluster baseado em Tabelas de Distância
SYKSOM	Rede de Kohonen
CLINT	Interpretação de *Clusters*
SCLASS	Árvore de Classificação não supervisionada
SPCA	Análise de Componentes Principais
SGCA	Análise Canônica
SDD	Descrição Discriminante
TREE	Árvore de Decisão
SDT	Árvore de Decisão Estratificada
SBTREE	Árvore de Decisão (Bayesiana)
SFDA	Análise Discriminante Fatorial

SREG	Análise de Regressão
SMLP	Perceptron Multicamadas

Tabela 1 – Módulos (métodos) do SODAS

Nesta dissertação e no estudo de caso do AondeNamoro.com tem-se como enfoque o estudo da estatística descritiva (métodos: DSTAT e VIEW) e os módulos de *Clustering (*Métodos: DIV, SCLUST, SYKSOM, SCLASS), tendo em vista atingir os objetivos acadêmicos e também para a empresa.

4 DATA MINING COM O SODAS

4.1 PREPARAÇÃO DE DADOS PARA O SODAS

O *Data Mining* do estudo de caso desta dissertação sobre o AondeNamoro.com, encontra-se dividido em 4 etapas: (WEISS, 1998).

- Preparação e Transformação dos Dados;
- Redução de Dados;
- Modelagem dos Dados e utilização do SODAS;
- Analises das Soluções;

A preparação dos dados é uma das etapas mais delicadas de todo o processo, pois através da mesma é norteada toda a preparação da análise de dados. Durante a fase de preparação, tem-se a oportunidade de conhecer melhor os dados, evento este necessário para qualquer análise multivariada, como é o caso do *Clustering*. Segundo Hair et all (1995) a Análise Multivariada demanda que se conheça rigorosamente os dados, tendo em vista que *outliers* ou dados não preenchidos podem ter efeitos significativos.

A grande difusão da tecnologia e das informações fez com que as empresas cada vez acumulassem mais dados, gerando diversas bases de dados. Sendo assim a preparação de dados é vital para analisar somente as informações relevantes, e não perder tempo e dinheiro com dados que não geram conhecimento. Segundo Pyle (1999), quando se está preparando os dados, está preparando o *Data Mining*, fazendo assim com que se consiga modelos mais confiáveis e rápidos.

No ponto de vista de Weiss (1998) a etapa mais crítica do processo de *Data Mining* é a preparação e transformação dos dados. Sendo algumas etapas da preparação de dados

realizadas durante a criação do *Data Warehouse*, entretanto normalmente são necessários ajustes para o *Data Mining*.

No estudo de caso apresentado nesta dissertação, a preparação dos dados para o *Data Warehouse* criado, já foi desenvolvida focando o *Data Mining* estudado nesta dissertação, tendo em vista que a empresa estudada ainda não possuía um *Data Warehouse*.

A preparação dos dados efetuada teve como conceito as Dez Regras de Ouro na Preparação de dados de Pyle (1999), sendo estas:

- Definição clara do problema e de seus benefícios;
- Especificar a solução desejada;
- Definir como a solução será utilizada;
- Entender o máximo possível do problema e dos dados;
- A modelagem deve ser baseada no problema;
- Refinar o modelo;
- Criar suposições;
- Fazer o modelo o mais simples possível;
- Definir a instabilidade do modelo, onde pequenas mudanças no input geram grandes mudanças no output;
- Definir as incertezas do modelo.

Conhecendo a importância da preparação de dados, dispensa-se uma boa quantidade de tempo, analisando qual seria a melhor forma de preparar o *Data Mining* e viabilizar a transformação da massa de dados em conhecimento sobre os clientes do AondeNamoro.com.

Nesta etapa do projeto foi necessário conhecer bem, tecnicamente, o funcionamento do SODAS, visando assim preparar as mudanças necessárias no banco de dados para que se pudesse ter o mesmo da melhor forma possível para a interação com o *software*.

Outro aspecto de grande importância desta etapa foi o conhecimento do banco de dados da empresa estudada, criando uma facilidade de identificar quais tipos de informações é possível extrair e assim ter uma noção melhor das possibilidades e restrições.

As variáveis que são utilizadas em nossa análise são baseadas nas respostas informadas pelos usuários da empresa estudada durante o cadastramento (ex: sexo, estado, estado civil, etc.) ou variáveis geradas através do histórico de navegação e interação com o site (ex: percentual de preenchimento do perfil do usuário, quantidade de mensagens enviadas, etc.).

Podem ser verificadas nas tabelas a seguir as variáveis preenchidas no cadastramento dos usuários. O processo de cadastramento é dividido em 4 etapas (Informações Principais, Como Sou, Detalhes e Meus Hábitos). (tabelas 2, 3, 4 e 5)

Informações Principais		
Variáveis	**Classificação da Variável**	**Exemplo**
Sexo	Qualitativa – Nominal – Dicotômica	(MASCULINO ou FEMININO)
Idade	Quantitativa – Discreta	(18,19, ...)
Estado	Qualitativa - Nominal	(RJ, ES, SP, MG, ...)
Cidade	Qualitativa – Nominal	(RIO DE JANEIRO, NITEROI, ...)
Bairro	Qualitativa – Nominal	(LEBLON, JARDINS, ...)

Intenção	Qualitativa – Nominal	(AMIZADE, RELACIONAMENTO SÉRIO, ...)
Estado civil	Qualitativa – Nominal	(SOLTEIRO, CASADO, VIUVO, ...)
Filhos	Quantitativa - Discreta	(0,1,2,3, ...)
Moradia: Moro com amigos	Qualitativa – Nominal	(SIM ou NÃO)
Formação	Qualitativa – Ordinal	(1 GRAU, 2 GRAU, SUPERIOR, ...)
Onde estuda(ou)	Qualitativa – Nominal	(IBMEC, FGV, ...)
Ramo profissional	Qualitativa – Nominal	(ADMINISTRAÇÃO, TURISMO, ...)
Ocupação atual	Qualitativa – Ordinal	(AUTONOMO, DESEMPREGADO, ...)
Opção sexual	Qualitativa - Nominal	(HETEROSSEXUAL, HOMOSSEXUAL, ...)
Faixa salarial	Qualitativa – Ordinal	(- de 1000, de 1000 a 2500, ...)
Com foto	Qualitativa – Nominal – Dicotômica	(SIM ou NÃO)

Tabela 2 – Variáveis da etapa "Informações Principais" do cadastro no *site*

Como Sou		
Variáveis	**Classificação da Variável**	**Exemplo**
Tipo físico	Qualitativa - Nominal	(MAGRO, ATLÉTICO, ACIMA DO PESO, ...)
Aparência	Qualitativa - Nominal	(BONITO, FEIO, ...)
Altura	Quantitativa - Continua	(1.40, 1.41, ...)
Peso	Quantitativa – Discreta	(40,41,....)
Cor dos olhos	Qualitativa – Nominal	(CASTALHOS, PRETO, VERDES, ...)
Cor do cabelo	Qualitativa – Nominal	(CASTANHOS, PRETOS, RUIVOS, ...)
Tipo do cabelo	Qualitativa – Nominal	(ONDULADO, LISO, ...)
Volume do cabelo	Qualitativa – Nominal	(CURTOS, MÉDIOS, LONGOS, ...)
Cor da pele	Qualitativa – Nominal	(BRANCA, NEGRA, ...)
Sinal/Característica	Qualitativa – Nominal	(CICATRIZ, BIGODE, ...)
Uso óculos	Qualitativa - Nominal	(SIM, NÃO ou USO LENTES)

Tabela 3 – Variáveis da etapa "Como Sou" do cadastro no AondeNamoro.com

Meus Hábitos e Detalhes		
Variáveis	**Classificação da Variável**	**Exemplo**
Exercício físico	Qualitativa - Ordinal	(NUNCA, 1 VEZ POR SEMANA, ...)
Esporte(s)	Qualitativa – Nominal	(FUTEBOL, BOX, VOLEI, ...)
Bebidas alcoólicas	Qualitativa - Ordinal	(SIM, NÃO E SOCIALMENTE)
Fumo	Qualitativa – Ordinal	(SIM, NÃO E SOCIALMENTE)
Costumo sair	Qualitativa – Ordinal	(NUNCA, 1 VEZ POR SEMANA, ...)
Quando saio vou a(o)	Qualitativa – Nominal	(CAMINHADAS, DISCOTECAS, ...)
Meu(s) hobbie(s)	Qualitativa – Nominal	(DESENHO, CINEMA, GASTRONOMIA, ...)
Gosta de Assistir Televisão	Qualitativa – Ordinal	(GOSTO, GOSTO MUITO, NÃO GOSTO, ...)
Tipos de programas de TV	Qualitativa – Nominal	(ESPORTE, NOVELA, TERROR, ...)
Leitura	Qualitativa – Ordinal	(GOSTO, GOSTO MUITO, NÃO GOSTO, ...)
Gênero de leitura	Qualitativa – Nominal	(POLICIAL, SUSPENSE, ...)
Gênero musical	Qualitativa – Nominal	(ROCK, BLUES, JAZZ,...)
Religião	Qualitativa – Nominal	(CATÓLICO, EVANGÉLICO, ...)
Quanto à prática religiosa	Qualitativa – Ordinal	(EVENTUAL, DEDICADO ou NÃO PRATICANTE)
Comida(s) preferida(s)	Qualitativa – Nominal	(PIZZA, CARNE, JAPONESA, ...)
Meus tipos preferidos de viagens	Qualitativa – Nominal	(PRAIA, CULTURAL, CASSINO, ...)
Meu(s) animal(is) de estimação	Qualitativa – Nominal	(CACHORRO, GATO, PEIXE, ...)
Gosto de vestir	Qualitativa - Nominal	(LARGADO, DESCONTRAÍDO, MODERNO, ...)

Tabela 4 – Variáveis da etapa "Meus Hábitos e Detalhes" do cadastro no *site*

Tendo em vista o trabalho de *Data Warehouse* efetuado, foram também geradas outras variáveis, que não são referentes ao cadastro dos usuários, mas um reflexo de toda sua navegação e interação com o AondeNamoro.com, estas variáveis foram criteriosamente criadas, visando identificar informações relevantes e estratégicas para empresa, sem ferir a privacidade dos usuários. Estas variáveis estão apresentadas tabela 5.

Variáveis geradas para o *Data Mining*		
Variáveis	**Classificação da Variável**	**Exemplo**
Percentual Cadastro Preenchido	Quantitativa – Discreta	(0 a 100)
Mensagens Enviadas	Quantitativa – Discreta	N
Mensagens Recebidas	Quantitativa – Discreta	N
Elogios[2] Efetuados	Quantitativa – Discreta	N
Elogios Recebidos	Quantitativa – Discreta	N
Perfis Visualizados[3]	Quantitativa – Discreta	N
Quantidade de vezes que seu perfil foi Visualizado[4]	Quantitativa – Discreta	N
Quantidade de logins efetuados	Quantitativa – Discreta	N
Impressões	Quantitativa – Discreta	N

Tabela 5 – Variáveis geradas exclusivamente para o DataMining

4.1.1 PRIVACIDADE NA UTILIZAÇÃO DOS DADOS DO AONDENAMORO.COM

Uma das maiores preocupações do AondeNamoro.com ao fornecer informações sobre o site, foi manter a privacidade e a segurança das informações dos usuários, tendo sido utilizado

[2] Elogios são mensagens pré-configuradas que um usuário do AondeNamoro pode enviar para outro.
[3] Será identificado nos gráficos e pelo SODAS como: total_perfis_visualizados
[4] Será identificado nos gráficos e pelo SODAS como: total_perfil_visualizado

diversos artifícios para que seja possível, em nenhuma hipótese, identificar de qual usuário pertencia cada informação.

Para manter a privacidade, os dados foram informados com o campo de identificação do usuário diferente do praticado pela empresa, sendo também utilizado um determinado índice para multiplicar as informações que nos foram passadas.

Isto demonstra que o AondeNamoro.com utiliza com bastante responsabilidade às informações de seus usuários, mantendo sua privacidade e respeitando os mesmos em todos os aspectos.

Outras informações e definições sobre Preparação de Dados podem ser obtidas em: (PYLE, 1999) e (WEISS, 1998).

4.1.2 TRATAMENTOS REALIZADOS

No conceito do banco de dados foram efetuadas poucas mudanças, entretanto com grande impacto no resultado, principalmente na agilidade em se trabalhar com a ferramenta de forma flexível.

A empresa estudada possui seu banco de dados relacional com informações dispersas em diversas tabelas, esta dispersão para um ambiente WEB é importante, pois dá mais agilidade e organização a informação. Entretanto, para trabalhar com o SODAS é mais interessante centralizar o máximo possível estas informações em poucas tabelas ou através de *views*, visando assim obter um desempenho mais rápido e também tornar a importação de dados mais simples. Sendo assim, para simplificarmos este processo, foram criadas *Views* no banco de dados, com as informações desejadas.

Uma *View*, segundo Kneipp e Albuquerque (2006), possui o funcionamento semelhante a uma tabela, quanto a recuperação e manipulação de dados (com algumas restrições), entretanto a mesma não possui a capacidade de armazenar os dados.

Durante a fase de tratamento de dados, é necessário efetuar toda preparação dos dados, para ter certeza que se está efetuando um *input* de forma correta. No caso do SODAS, a preparação dos dados também é um momento onde deve ser otimizado o banco afim de também serem aproveitadas todas as diversas funcionalidades oferecidas pelo *software*.

Pensando nisto foi criada uma tabela "tbTaxo", destinada a armazenar informações para serem utilizadas unicamente pelo SODAS, para criar suas taxonomias, que é explicada de forma mais completa na demonstração prática da importação (Apêndice A).

4.1.3 IDENTIFICAÇÃO DE *OUTLIERS*

De acordo com Han e Kamber (2001), *outliers* são dados que não se encontram no padrão de "comportamento" do modelo de dados. Os *outliers*, normalmente são descartados pelo *Data Mining* como exceções. A análise de *outliers* para muitas empresas é de alta relevância, como por exemplo, para operadoras de cartão de crédito analisarem fraudes.

Para What (2006) *outliers* são elementos (observações) que possuem um valor anormal quando comparado com outros elementos da amostra.

Nesta etapa, busca-se identificar registros no banco de dados que possuíssem informações não corretas, ou por erro de digitação ou por erro do próprio sistema.

No estudo de caso apresentado, a identificação de *outliers* foi facilitada, pois por tratar-se de um sistema completamente via WEB e com perguntas fechadas, os usuários não teriam teoricamente como gerar informações incorretas.

Porém, apesar de toda proteção do sistema, foi identificado um tipo de *Outlier* no campo referente à idade dos usuários, que devido a uma falha no sistema da empresa estudada alguns poucos usuários estavam com idades completamente fora do padrão. Sendo assim foram excluídos da análise os registros de usuários com menos de 18 anos (não é permitido o cadastramento de usuários menores de idade, entretanto existiam usuários com idade inferior a zero anos) e usuários com mais de 90 anos.

Existem quatro tipos de outliers, sendo estes: (HAIR et al, 1995)

- Erros de procedimento;
- Eventos extraordinários;
- Evento extraordinário os quais a analise não conseguiria "explicar";
- Eventos (observações) nas quais se encontram dentro de uma *range* e que são únicas em sua combinação de valores.

Os *outliers* devem ser sempre analisados, para ser avaliada a melhor forma de tratar estes dados. No estudo de caso, por tratar-se de um *outlier* de "Erro de procedimento", optou-se por excluí-los, tendo em vista que os mesmos compõem somente 177 registros e mantê-los na análise iria enviesar todo o resultado da mesma.

4.2 IMPORTAÇÃO BANCO DE DADOS

Um dos pontos críticos de sucesso para a utilização de uma ferramenta de *Data Mining* é a possibilidade da mesma poder interagir com o SGBD (Sistema de Gerenciamento de Banco de Dados) utilizado pela empresa. O *Software* SODAS tem a capacidade de interagir com diversos tipos de banco de dados, como por exemplo: Access, Mysql, Paradox, CVS, entre outros. O SODAS possui a grande flexibilidade de poder se conectar com qualquer SGBD que tenha suporte a conexão através de ODBC (*Open Data Base Connectivity*).

Segundo Martins (2006) ODBC (*Open Database Connectivity*) é uma interface para acesso a dados, desenvolvida por grupos de padronização como *SQL Access Group*. O ODBC permite que uma determinada aplicação interaja com uma grande variedade de Sistemas de Gerenciamento de Banco de Dados (SGBD) podendo efetuar diversas funções, sendo as principais: conectar e desconectar fontes de dados, preparar e executar comandos, processar erros e processar transações.

No estudo de caso a empresa utiliza um banco de dados livre chamado *MYSQL*, que atualmente é um dos bancos de dados mais utilizados no mundo por *sites* de internet. O SODAS possui completa sinergia para trabalhar com o MYSQL, tendo em vista que o mesmo aceita conexão via ODBC.

MYSQL é um sistema de gerenciamento de banco de dados relacional, baseado em SQL (*Structured Query Language*), que teve sua primeira versão lançada em janeiro de 1998. Este banco de dados é principalmente utilizado por empresas de internet, tendo em vista que o mesmo é um *software* livre e com código fonte aberto, o que criou uma grande vantagem

frente aos outros sistemas proprietários como, por exemplo: Oracle, IBM, Informix entre outros. (MYSQL, 2006).

Esta sinergia entre a combinação de MYSQL + SODAS chamou atenção da empresa do estudo de caso, devido a viabilidade de se utilizar um banco de dados livre e também uma ferramenta livre para mineração de dados, o que faz com que a mesma seja uma solução com um excelente custo benefício para empresa.

Um dos primeiros passos para que o SODAS realize o *Data Mining* de forma correta é que se tenha uma boa base de dados, seguindo os conceitos de um banco de dados relacional. Pode-se visualizar, através da figura 4, a representação de algumas tabelas do banco de dados do AondeNamoro.com, demonstrando que o mesmo utiliza os conceitos de relacionamento.

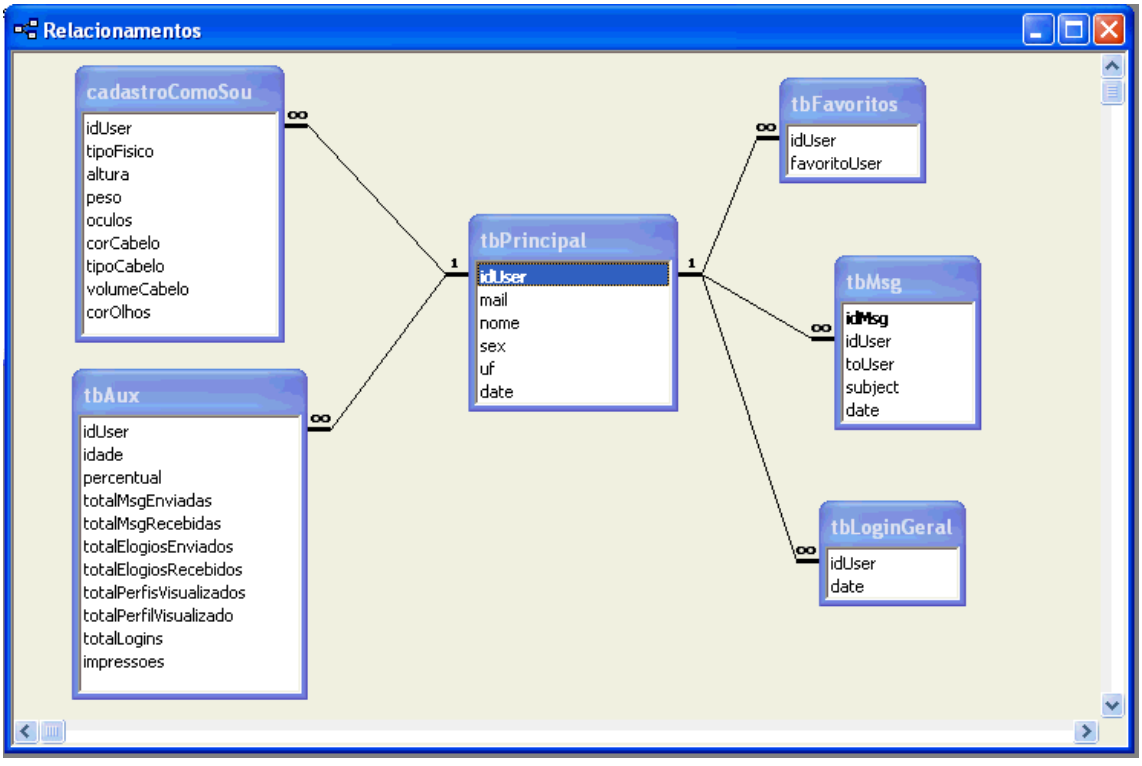


Figura 4 – Exemplo de parte do banco relacional do AondeNamoro.com

O SODAS possui uma ferramenta específica chamada DB2SO, que é instalada juntamente com o *software*, sendo esta responsável por efetuar toda a importação de dados e também a manipulação destes dados, como, por exemplo, para criação de taxonomias e regras de dependência.

No estudo de caso foi efetuada a importação através do DB2SO, onde foi gerado um arquivo "*.sds" para utilização pelo SODAS, com as informações de 4 variáveis qualitativas e 10 variáveis quantitativas da base de dados do AondeNamoro.com. Também foi criada uma taxonomia para a variável "estado", onde os estados foram divididos por regiões. O *output* da importação pode ser verificado na figura 5. Maiores dados e informações sobre o procedimento de importação de dados para o SODAS podem ser verificados no tutorial passo a passo disponível no Apêndice A.

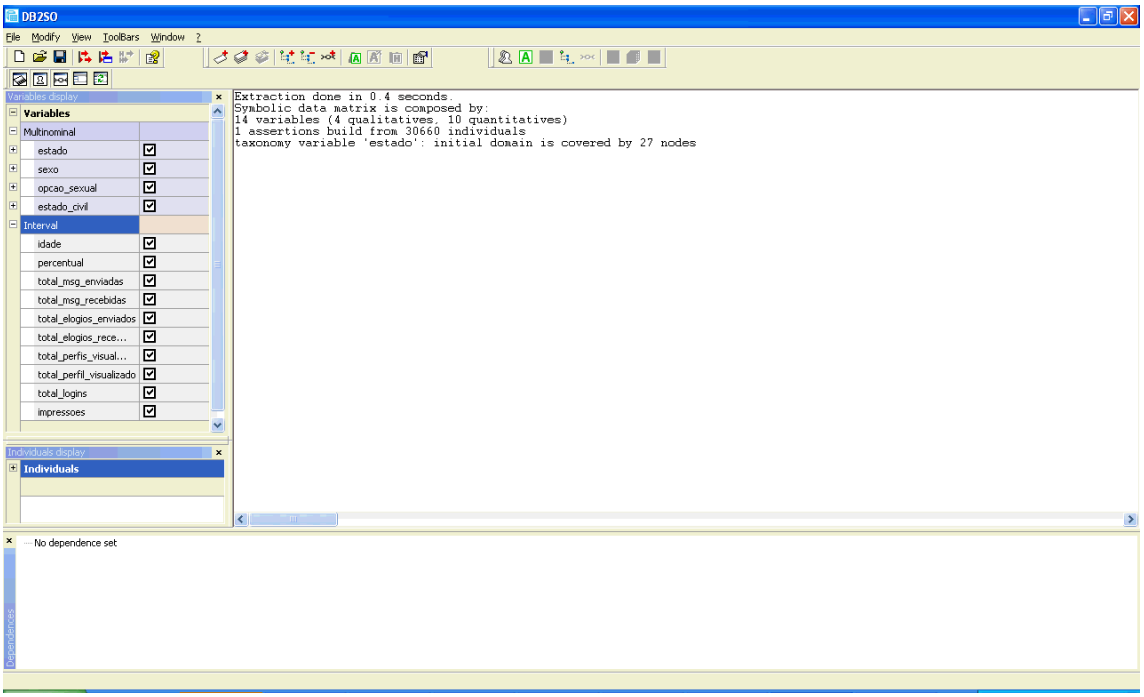


Figura 5 – *Output* importação para o SODAS

4.3 MINERAÇÃO DOS DADOS

4.3.1 ESTATÍSTICA DESCRITIVA

Nesta seção tem-se uma visão mais prática do funcionamento do SODAS para *data mining*, onde é realizada a estatística descritiva dos dados do AondeNamoro.com, utilizando os módulos VIEW e DSTAT do SODAS. O tutorial passo a passo da utilização destes métodos esta disponível no Apêndice B.

O método *View*, como seu próprio nome ressalta, tem como funcionalidade principal demonstrar através de diversos gráficos em 2d e 3d informações básicas sobre os dados simbólicos e as variáveis utilizadas no modelo.

O método DSTAT é mais completo que o View, oferecendo mais informações (histogramas, médias, mínimo, máximo, entre outros), e possui gráficos mais específicos.

Inicialmente são gerados gráficos em 3d utilizando o método View, para se ter uma visão melhor das informações disponíveis e assim começar a entender melhor os dados e obter uma concepção mais formada de quais informações pode-se extrair. É utilizado inicialmente o sexo como Objeto Simbólico, sendo assim o gráfico fica dividido entre “MASCULINO” e “FEMININO”. (Fig. 6)

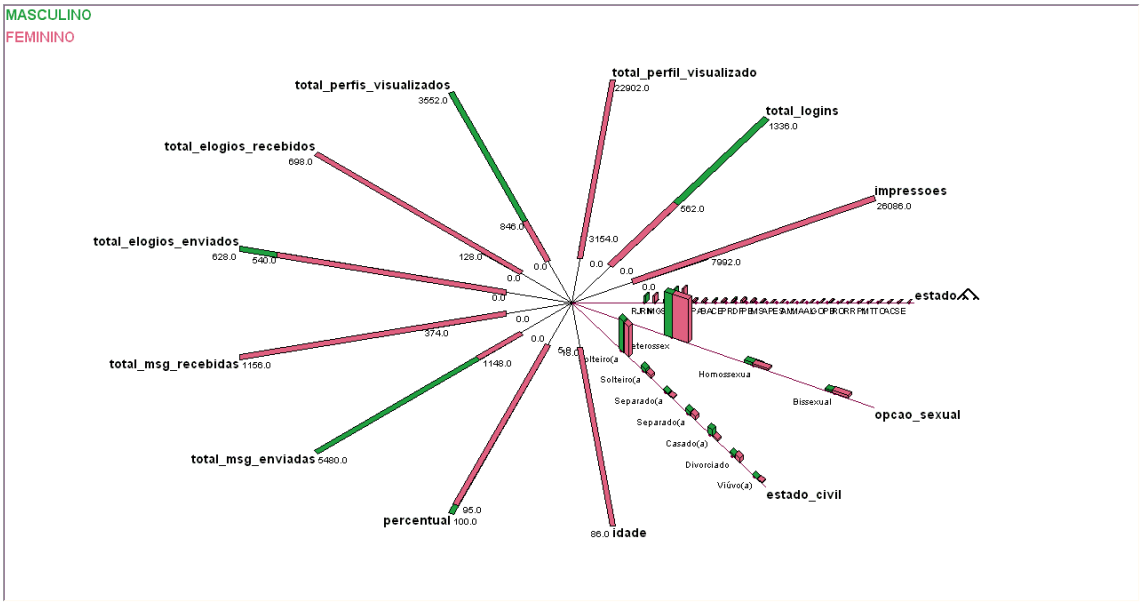


Figura 6 – View todas variáveis, base "Sexo" (método VIEW)

Através da interpretação da figura 6, já se consegue extrair algumas informações básicas. Pode-se verificar que os homens em geral efetuam mais *logins* (momento no qual o usuário acessa o AondeNamoro.com com seu "apelido" e "senha") do que as mulheres, entretanto as mulheres têm um potencial de gerar mais impressões do que os homens.

Pode-se também verificar através da figura, que os homens preenchem normalmente o perfil mais do que as mulheres, o que pode mostrar que os mesmos dedicam mais tempo aos seus dados, provavelmente devido a maior concorrência para os homens, que estão em grande maioria no site. Quanto à idade, pode-se verificar que os homens e as mulheres estão em igualdade em limite de idade.

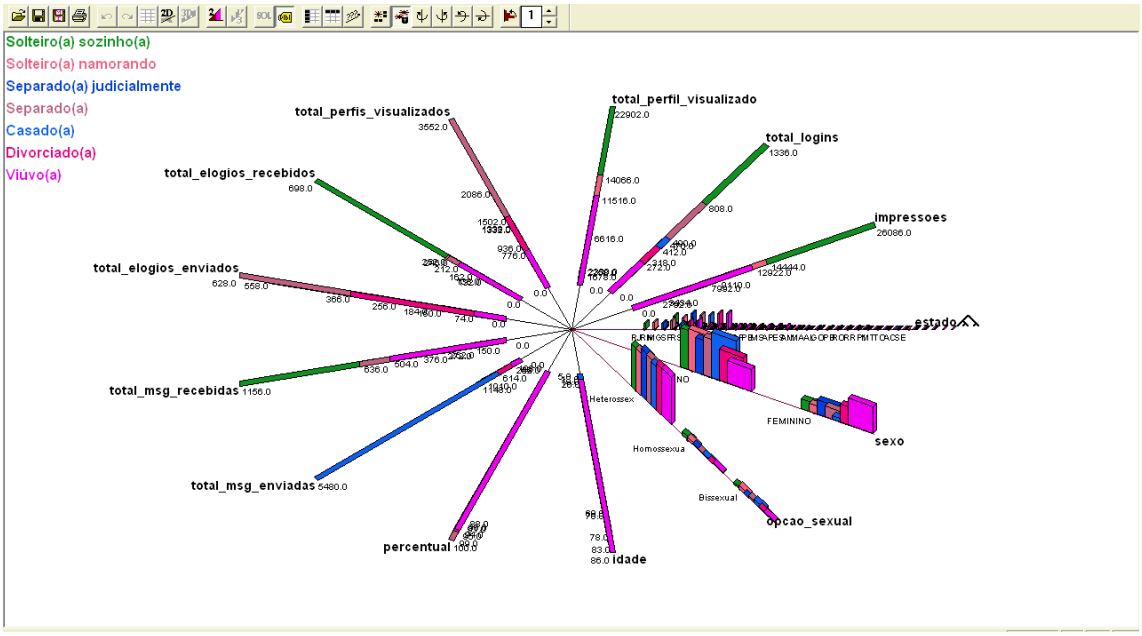


Figura 7 – View todas variáveis, base "Estado Civil" (método VIEW)

Gerando o gráfico do método *View*, utilizando a variável "Estado Civil" como base (Fig. 7), pode-se verificar que os usuários com perfil de "Solteiro sozinho(a)", ou seja, usuários que não possuam compromisso, possuem aparentemente o melhor perfil, pois são os que possuem o maior limite de impressões, de quantidade de logins efetuados no *site*, os que possuem seus perfis mais vezes visualizados e são os que mais recebem elogios e mensagens.

Ainda analisando a figura 7, identifica-se que os usuários com estado civil de "separado", são os que possuem seus perfis mais completos (variável "percentual") e também são os que visualizam mais perfis e costumam a enviar mais elogios. Um fato curioso é que os usuários "separados" apesar de serem os que visualizam mais perfis e serem os que enviam mais elogios, não enviam muitas mensagens. Isto demonstra uma certa timidez dos usuários "separados", pois é muito mais simples enviar um elogio, tendo em vista que as mensagens são pré-configuradas, do que enviar uma mensagem.

Através da análise da figura 8, pode-se perceber que 92% dos usuários casados são do sexo “Masculino”, contra somente 8% do sexo “Feminino”, isto demonstra que possivelmente as mulheres são mais fiéis aos seus parceiros, ou então preenchem erroneamente esta variável. Seria interessante para empresa criar alguma outra variável em seu cadastro que pudesse validar a resposta utilizada na variável “Estado Civil”.

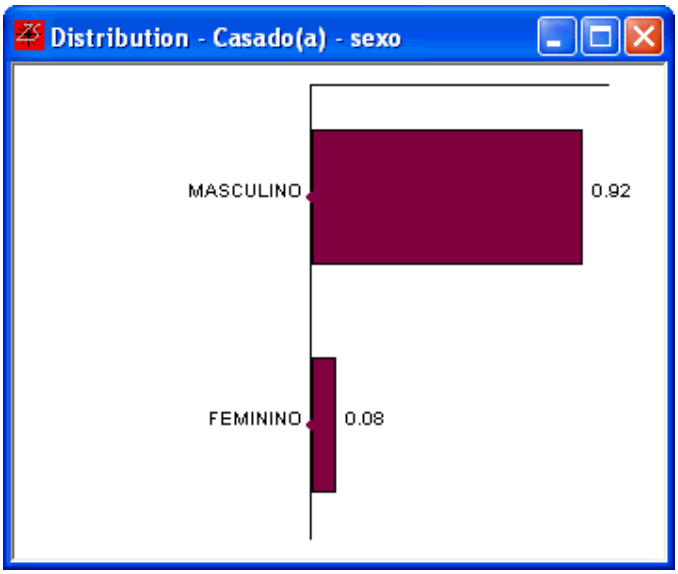


Figura 8 – Distribuição do Sexo entre usuários Casados(as).

Um fato curioso que se pode verificar através da figura 9, é que no estado civil de “Viúvo(a)” 55% dos usuários são do sexo “Feminino”, sendo interessante para empresa verificar o porque deste fator, para compreender melhor as necessidades deste tipo de usuário.

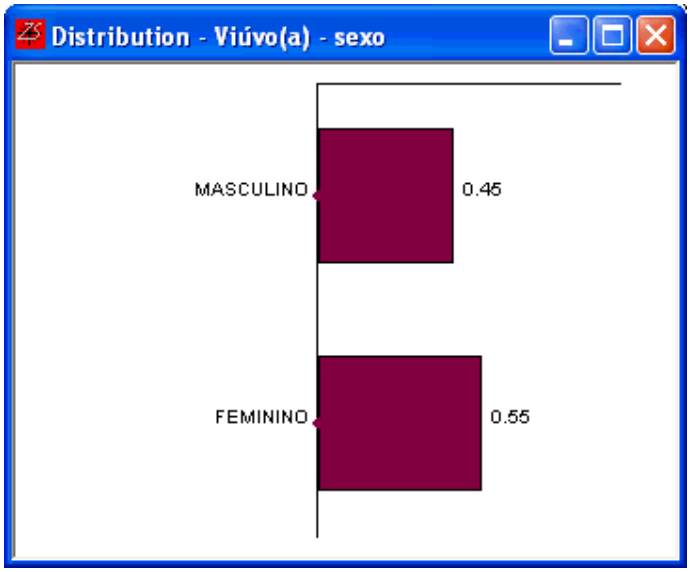


Figura 9 – Distribuição do Sexo entre usuários Viúvos(as).

As análises a seguir são obtidas trabalhando com o método DSTAT, onde se tem a possibilidade de visualizar as informações de forma mais detalhada e clara e também a possibilidade de confirmar as informações anteriores.

Pode ser percebido através da figura 10, que a grande maioria dos usuários do AondeNamoro.com são do sexo “Masculino” (aproximadamente 80%). Esta informação é simples, porém bem relevante, pois aponta um possível problema para “criação” de relacionamentos, tendo em vista que há uma quantidade excessiva de homens e poucas mulheres, gerando uma grande dificuldade para que os homens localizem o seu par feminino ideal, causando assim uma possível desistência da utilização do serviço.

Uma das recomendações importantes para o AondeNamoro.com é que consiga ampliar sua base de usuários femininos, tendo em vista que a maioria dos usuários da internet são mulheres o que demonstra que existe algo de errado na comunicação para as mesmas e também a oportunidade de um grande público ainda não muito explorado. (MAIORIA, 2006)

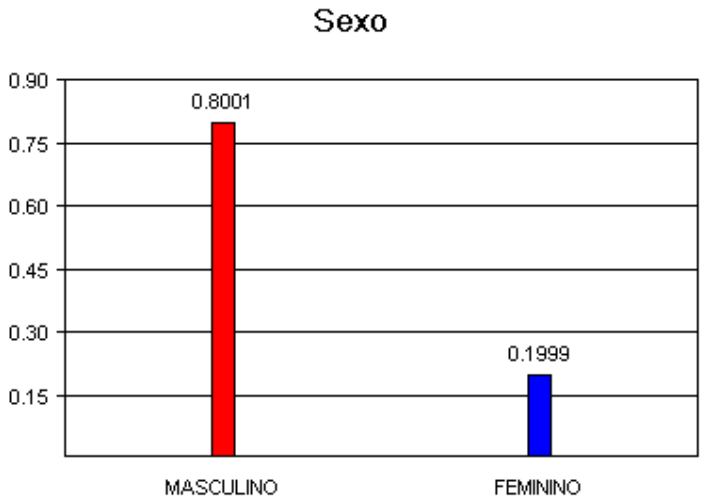


Figura 10 – Gráfico Sexo (método DSTAT)

A participação de cada estado civil na base de dados do AondeNamoro.com (Fig. 11) é uma informação extremamente relevante, pois a auxilia a identificar necessidades importantes para o usuário, principalmente quanto ao que ele espera do site, como, por exemplo: amizade, relacionamento casual, relacionamento sério, entre outros.

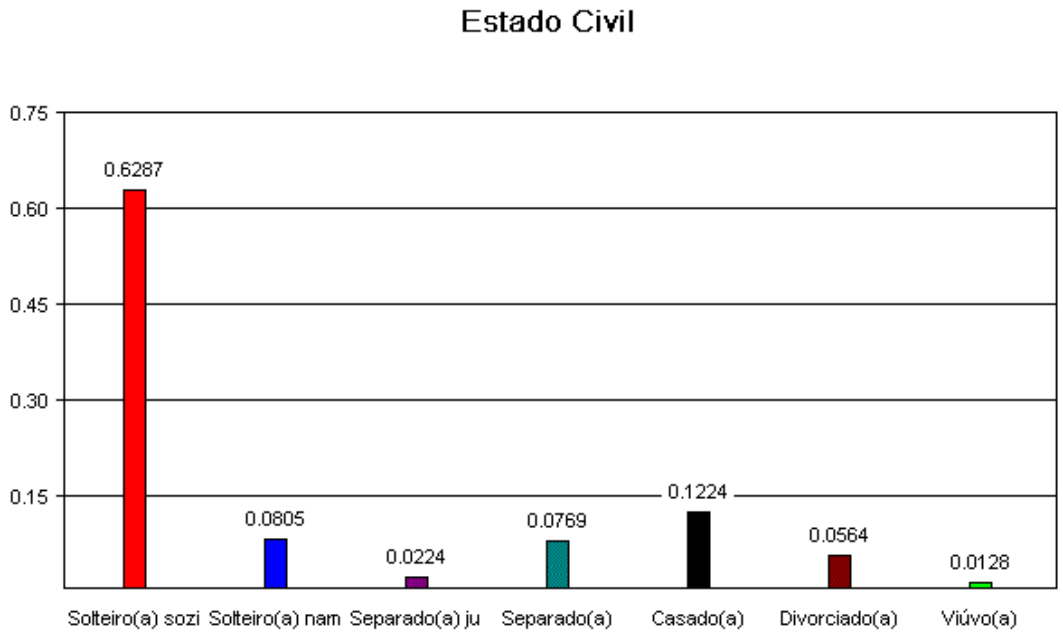


Figura 11 – Gráfico Estado Civil (método DSTAT)

A opção sexual (Fig. 12) representa a distribuição da preferência sexual de nossos usuários. Este dado é muito importante para o site, tendo em vista que podem ser criadas campanhas especificas para cada preferência, tendo em vista que cada uma possui necessidades diferentes. Podem ser realizados trabalhos maiores com *Data mining*, para se compreender as especificidades de cada opção sexual e o comportamento destes usuários no *site*. Entretanto, a pedido do AondeNamoro.com este tipo de análise não é realizado nesta dissertação, visando proteger a privacidade dos usuários.

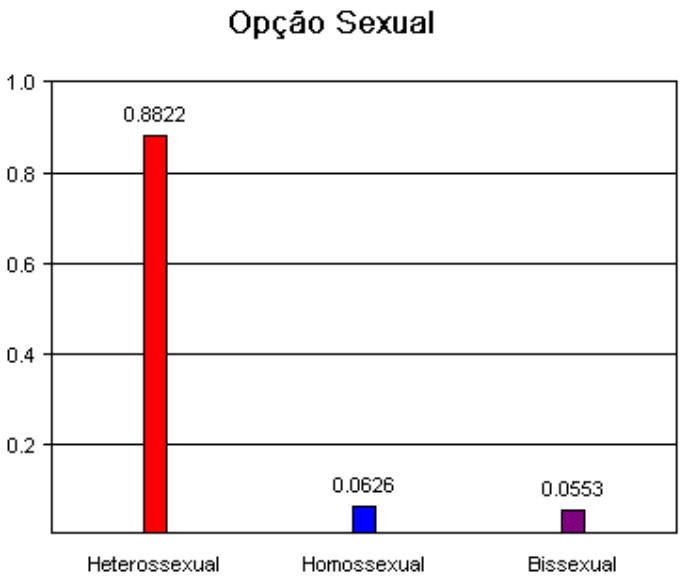


Figura 12 – Gráfico Opção Sexual (método DSTAT)

O gráfico de Impressões (Fig. 13) abrange a informação de quantas impressões cada tipo de usuário gerou para o AondeNamoro.com. Esta informação é relevante para que o AondeNamoro.com possa ter uma estimativa de receita que aquele usuário pode gerar para empresa, uma vez que no novo formato de receita que a empresa adotará, a mesma será remunerada de acordo com a quantidade de páginas vistas durante a navegação do usuário. Pode-se verificar que apesar do público masculino ser a maioria no site, as mulheres são as que geram mais impressões, conseqüentemente faturamento, para empresa. Este dado foi uma grande surpresa no estudo, e existem neste momento duas hipóteses para este evento. A

primeira devido ao fato de as mulheres serem as que recebem mais elogios e mensagens, o que estimula que as mesmas acessem mais o serviço. A segunda hipótese é que apesar de serem a minoria, as mulheres devem ser usuárias mais ativas no site, enquanto os homens o utilizam com menos freqüência.

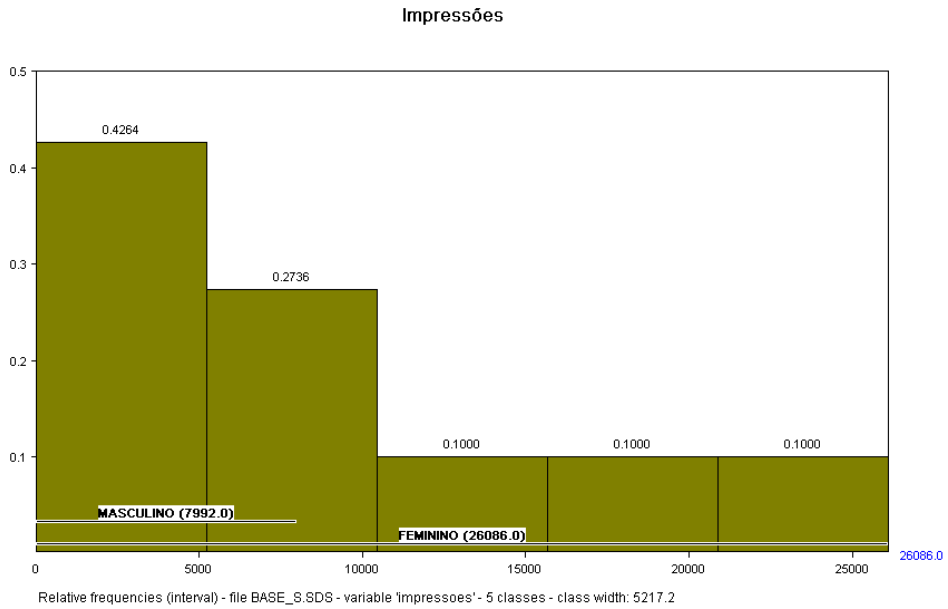


Figura 13 – Gráfico Impressões Geradas (método DSTAT)

As mulheres apesar de gerarem mais tráfego para o AondeNamoro.com, não são as que geram maior quantidade de logins (Fig. 14). Este fato provavelmente ocorre por que as mulheres devem permanecer “logadas” durante mais tempo no serviço do que os homens, o que faz com que sua quantidade de *logins* seja reduzida e também devido ao fato dos homens serem os que mais enviam mensagens e elogios, operações que somente podem ser realizadas com o usuário “logado” no AondeNamoro.com.

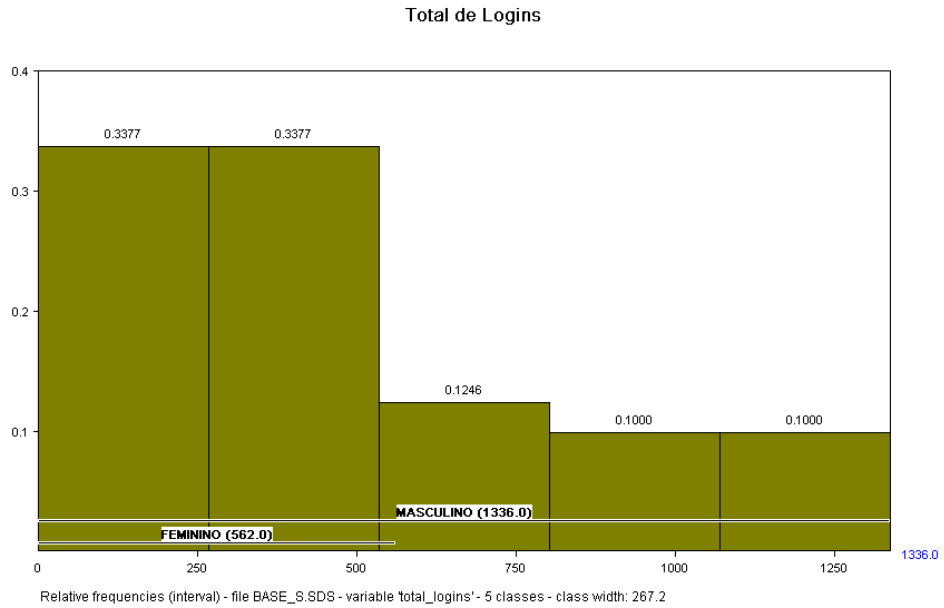


Figura 14 – Gráfico Quantidade de Logins realizados (método DSTAT)

No gráfico da figura 15, é visualizado um gráfico "Biplot" utilizando as variáveis Impressões e Total de Logins, para que se possa ter uma visualização melhor do fenômeno demonstrado inicialmente pelas figuras 13 e 14.

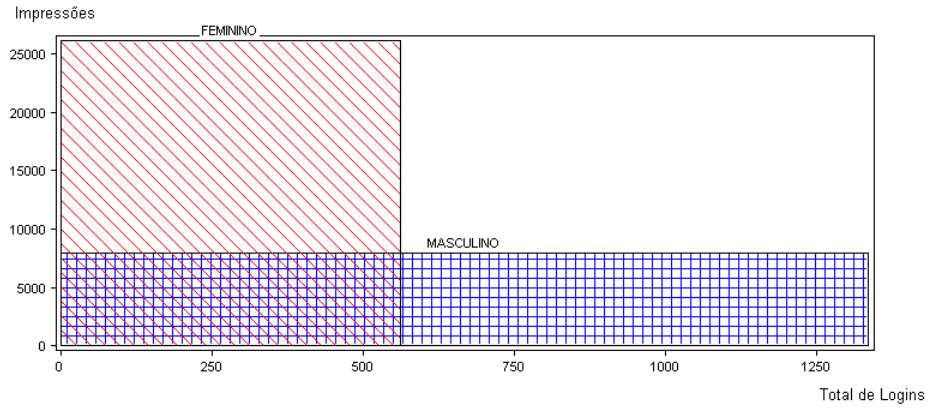


Figura 15 – Gráfico Biplot (Impressões X Total de Logins) (método DSTAT)

Pode-se verificar também, através do próximo gráfico biplot (Fig. 16), onde ocorre o cruzamento das variáveis “Total Mensagens Recebidas” e “Total Mensagens Enviadas” uma outra explicação para o fenômeno, tendo em vista que as mulheres recebem cerca de 5 vezes mais mensagens do que os homens e verificar uma mensagem gera mais impressões do que enviar uma, sendo este também um forte indício para o motivo do público feminino gerar mais impressões do que os homens.

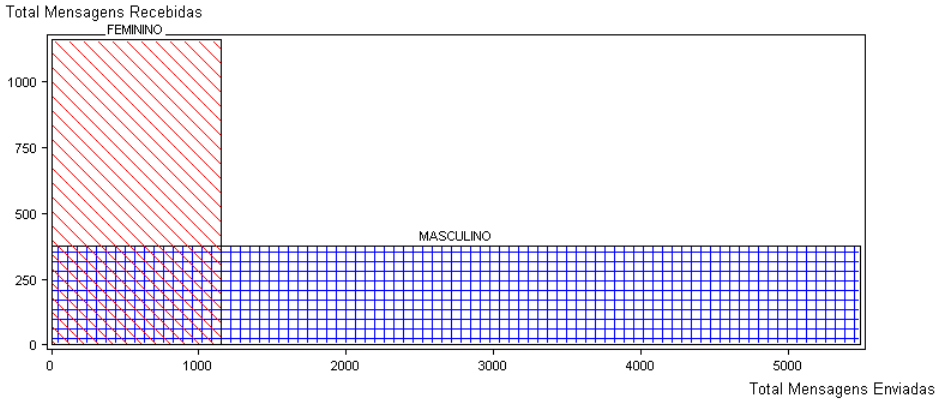


Figura 16 – Gráfico Biplot (Total Mensagens Recebidas X Total Mensagens Enviadas) (método DSTAT)

Na figura 17, é ilustrado um gráfico biplot, onde utiliza como base o “Estado Civil” e são avaliado as variáveis “impressões” e “total_perfil_visualizado”. Verifica-se assim que os usuários de estado civil “Solteiro Sozinho” geram mais impressões e são os que têm seus perfis mais visualizados, sendo um publico altamente lucrativo, em escala decrescente de atratividade para empresa, temos os seguintes estados civis: “Solteiro Sozinho”, “Solteiro Namorando”, “Viúvo(a)”, “Separado(a)”, “Casado”, “Divorciado” e por fim “Separado Judicialmente”.

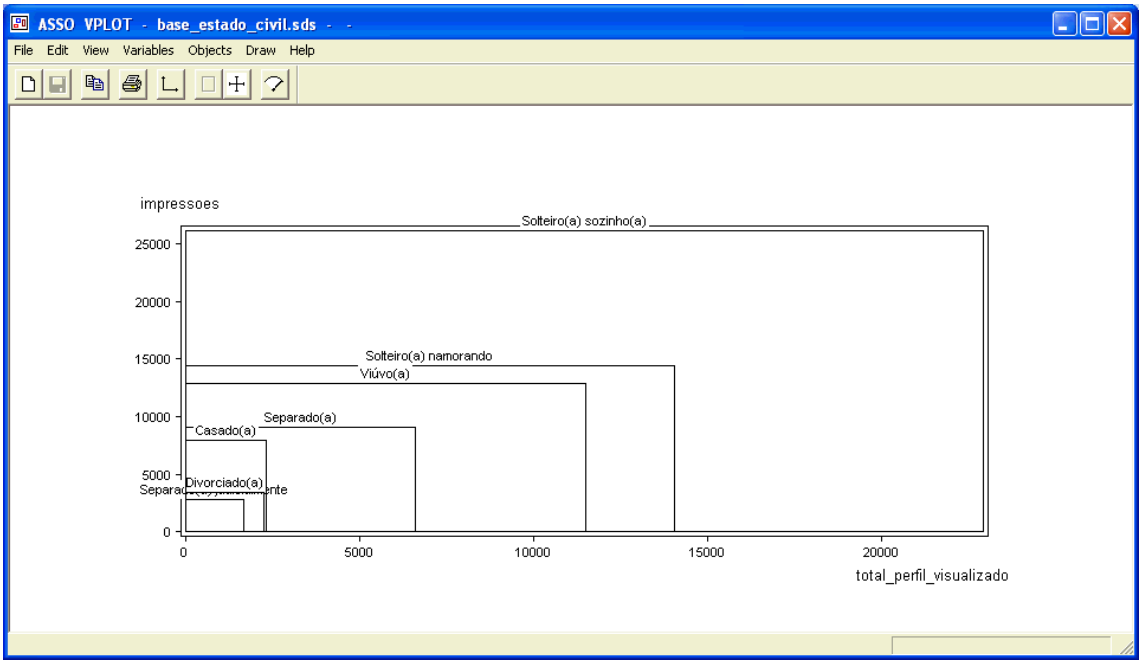


Figura 17 – Gráfico Biplot (Impressões X Perfil Visualizado) .

Buscando entender melhor o fenômeno e conhecer melhor a base de dados, foi utilizado o parâmetro “Numeric *and symbolic characteristics”* do método DSTAT, onde se verifica que existe uma correlação de 0,896 entre as variáveis “total_msg_recebidas” e a variável “impressões”, quando se utiliza a variável “estado” como sendo o objeto simbólico. (Fig. 18)

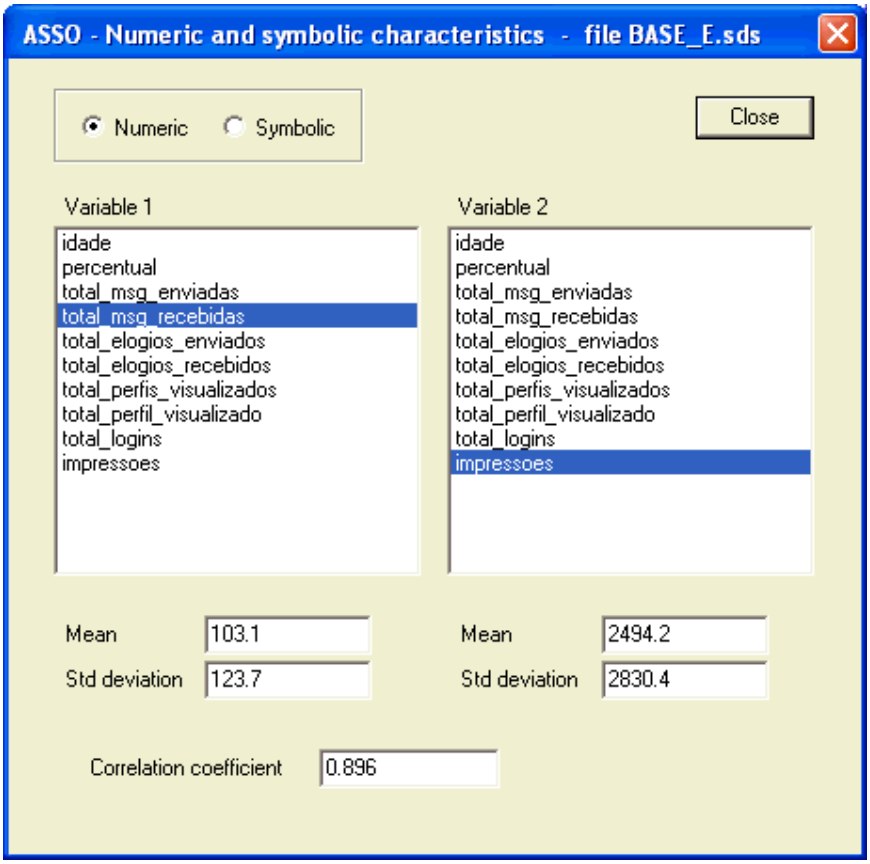


Figura 18 – *"Numeric and symbolic characteristics" ("total_msg_recebidas x "impressoes")*

Verifica-se também utilizando o parâmetro "*Numeric and symbolic characteristics*", que existe uma forte correlação (0,855) entre as variáveis "total_msg_enviadas" e "total_msg_recebidas", o que leva a crer que usuários que enviam mensagens recebem muitas mensagens, sendo assim usuários pouco participativos recebem poucas mensagens.(Fig. 19)

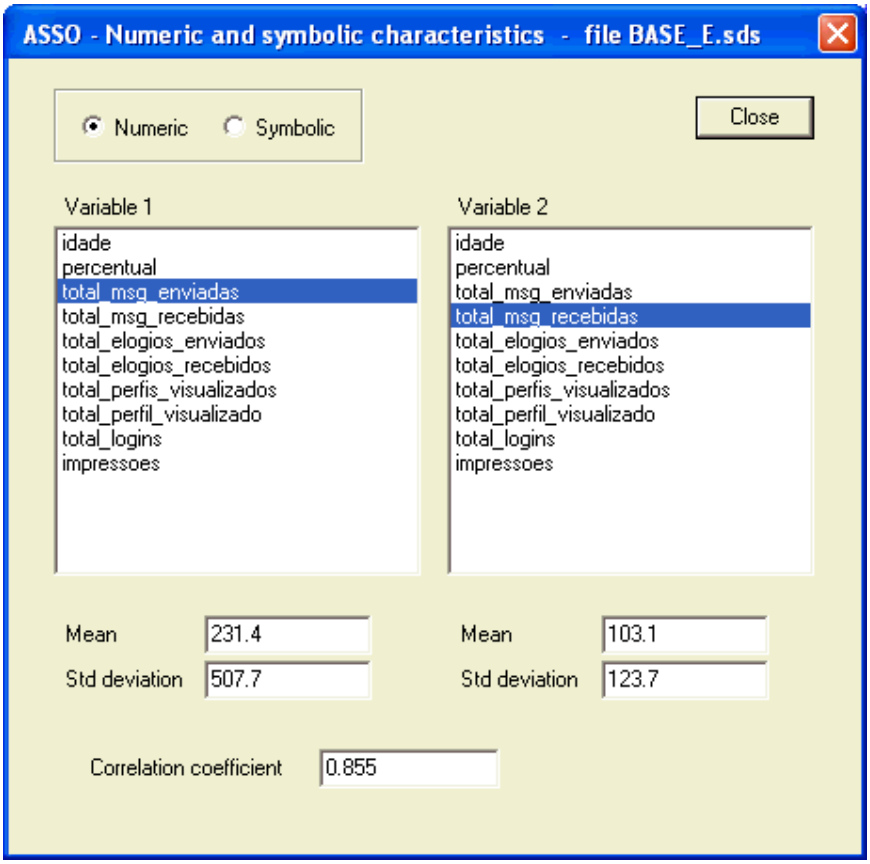


Figura 19 – *"Numeric and symbolic characteristics" ("total_msg_enviadas x "total_msg_recebidas")*

4.3.2 CLUSTERING

Após realizar o estudo da estatística descritiva do AondeNamoro.com, iniciou-se a utilização dos métodos de *Clustering* do SODAS.

O SODAS por ser um *software* de *Data Mining*, possui diversos métodos relacionados a *Clustering*, os quais nesta etapa estaremos utilizando os módulos: DIV, SCLUST, SYKSOM e SCLASS.

O método DIV efetua uma Classificação de Dissimilaridades, sendo o mesmo muito interessante para gerarmos *clusters*. Segundo Lechevallier e Chavent (2003) o método

"*divisive classification"* trabalha com *Clustering* hierárquico, onde todos os objetos iniciam em um único *cluster* e ocorrem sucessivas divisões gerando os outros *clusters.* O algoritmo encerra os *clusters* em (K – 1) divisões, onde K é o número de *clusters* definido pelo usuário nos parâmetros de configuração do DIV.

O SCLUST é um método de *Clustering* dinâmico utilizado pelo SODAS, onde de acordo com Carvalho e Lechevallier (2003), é utilizado para dividir *n* objetos simbólicos de *p*-dimensões em *m clusters* homogêneos.

O método SYKSOM (Rede de Kohonen) semelhantemente ao SCLUST também é utilizado para dividir *n* objetos simbólicos de *p*-dimensões em *m clusters* homogêneos, entretanto no SYKSOM é utilizada uma analogia a construção de mapas de Kohonen. Este método também se caracteriza por ter uma ênfase na demonstração gráfica do resultado. (SYKSOM, 2003)

O método SCLASS é considerado uma "Árvore" de Classificação não supervisionada, onde semelhantemente ao DIV, também trabalha com um método de divisão de *cluster*, onde todos os objetos iniciam em um único *cluster* e ocorrem sucessivas divisões, e cada *cluster* se transforma em dois menores até que seja atingida a regra de parada. Entretanto, o SCLASS apresenta características em seu algoritmo que o diferencia do DIV.

Inicialmente é utilizado o método DIV, onde são selecionadas todas as variáveis do tipo Interval para análise no SODAS. Pode-se verificar através da figura 20, a formação dos *clusters* dos objetos simbólicos (neste caso "estado") e também as variáveis escolhidas como variáveis de corte pelo SODAS. O *software* utiliza variáveis diferentes para cada divisão, onde neste primeiro momento existe uma coerência na divisão efetuada pelo método DIV, tendo em vista que separou os estados em *clusters*, onde os mesmos realmente possuem

semelhanças (ex: *cluster* RJ, MG e RS). Através deste método a empresa pode-se identificar semelhanças de comportamento no *site* por parte de usuários de estados diferentes, o que poderá reduzir custos de criação de peças de divulgação e facilitará a criação nichos no *site.*

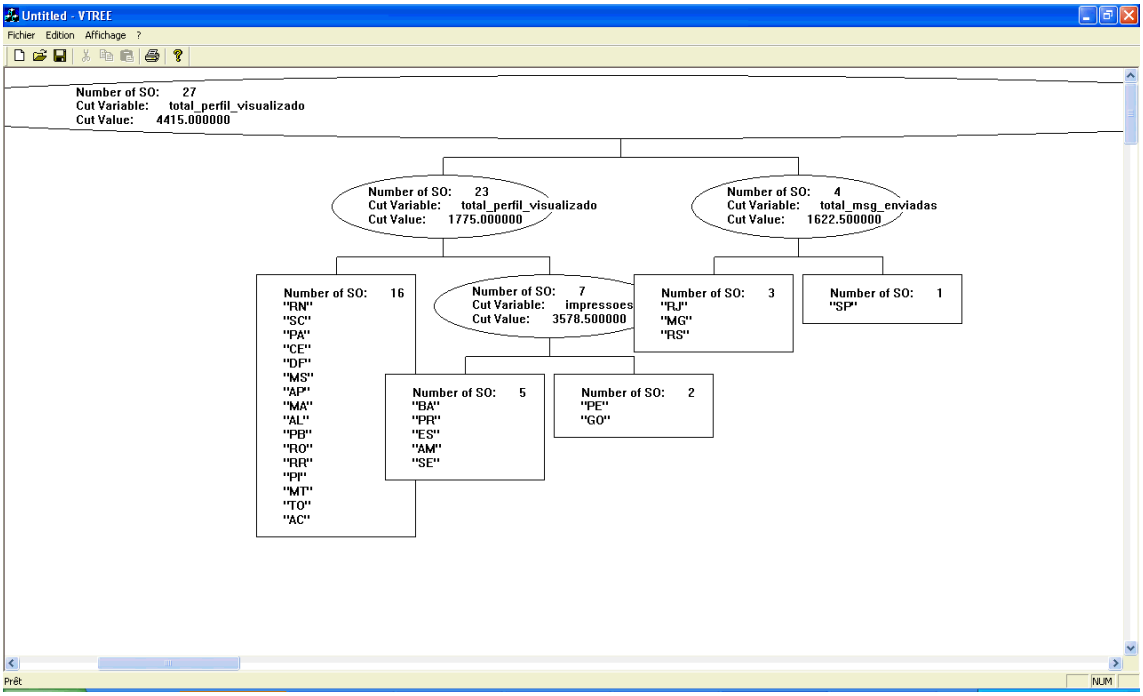


Figura 20 – Árvore com *clusters* (método DIV)

O método DIV é muito interessante, pois permite gerar *clusters* a partir do *input* de quantas variáveis se desejar, onde no exemplo anterior o SODAS escolheu quais variáveis utilizar como padrão para o corte. Entretanto, pode-se forçar para ser utilizada sempre somente uma variável. Sendo assim, é possível "forçar" o método DIV do SODAS a somente trabalhar com a variável "impressões" e assim conhecer melhor os *clusters* de estados para esta variável e identificar os estados mais lucrativos para empresa. (Fig. 21)

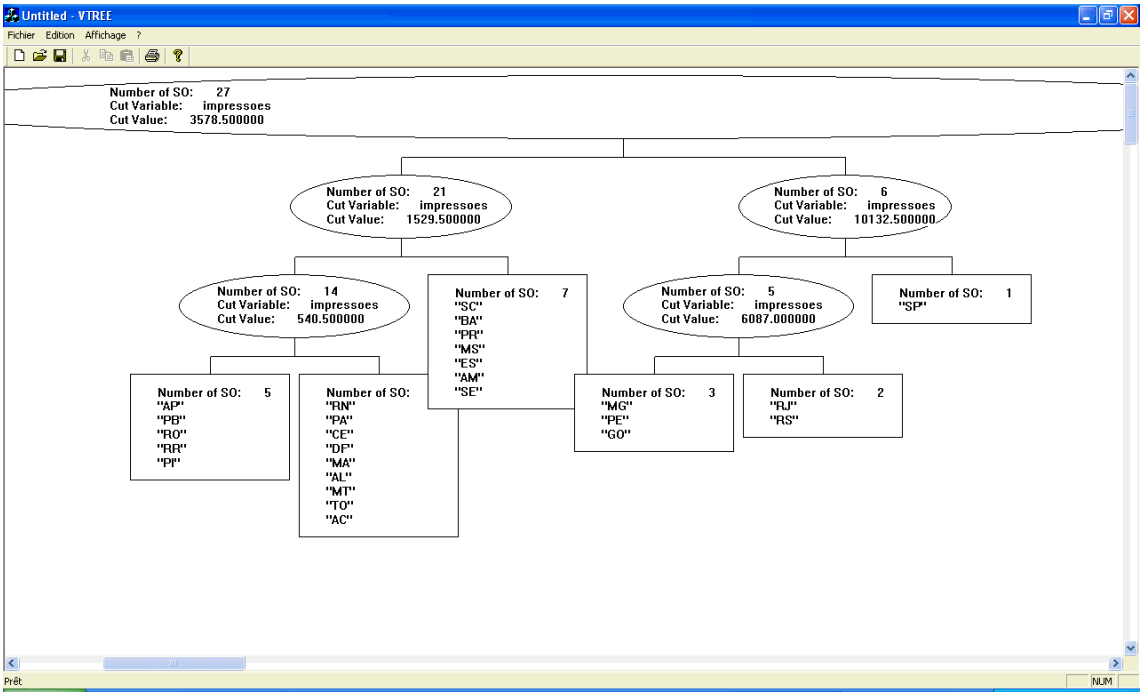


Figura 21 – Árvore com 6 *clusters* utilizando variável impressões (método DIV)

Na figura 21, pode-se verificar que foram gerados 6 *clusters* de estados baseado na variável "impressões". A árvore gerada nos mostra que o *cluster* de melhor resultado é o composto somente pelo estado SP, que possui em seu limite mais de 10.132 impressões, seguido pelos *cluster* dos estados RJ e RS, que possuem mais de 6.087 impressões. Pode-se verificar mais facilmente esta diferença analisando os arquivos de *logs*. (Fig. 22 e 23).

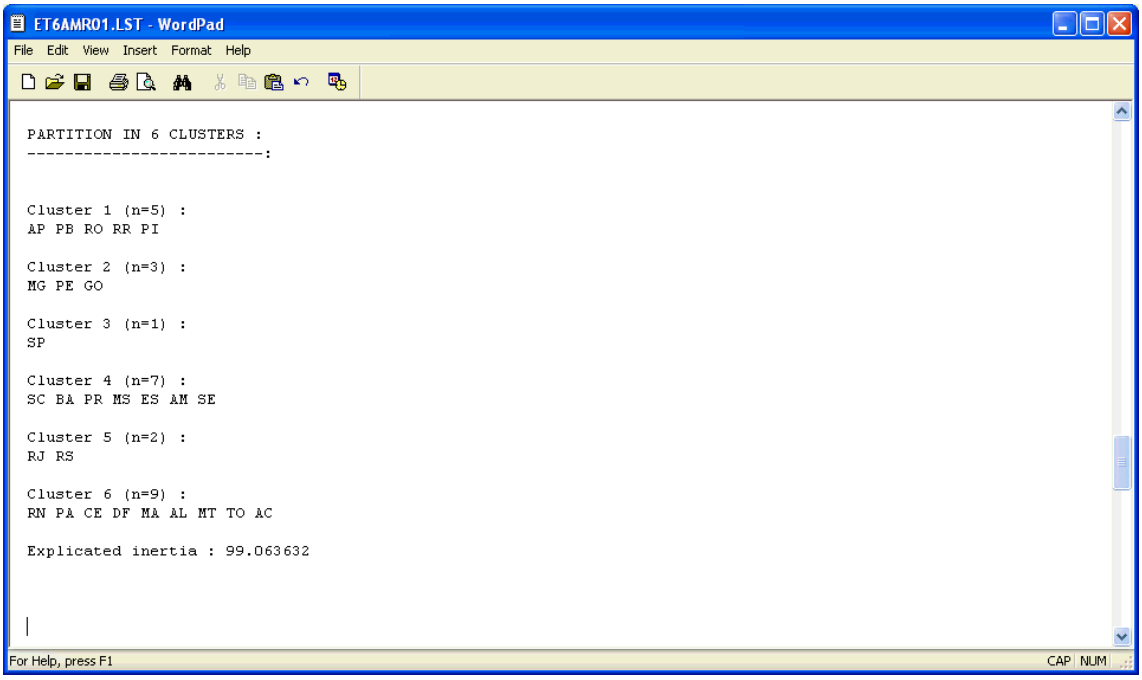


Figura 22 – *Log* dos dados da árvore com 6 *clusters* utilizando variável impressões (método DIV)

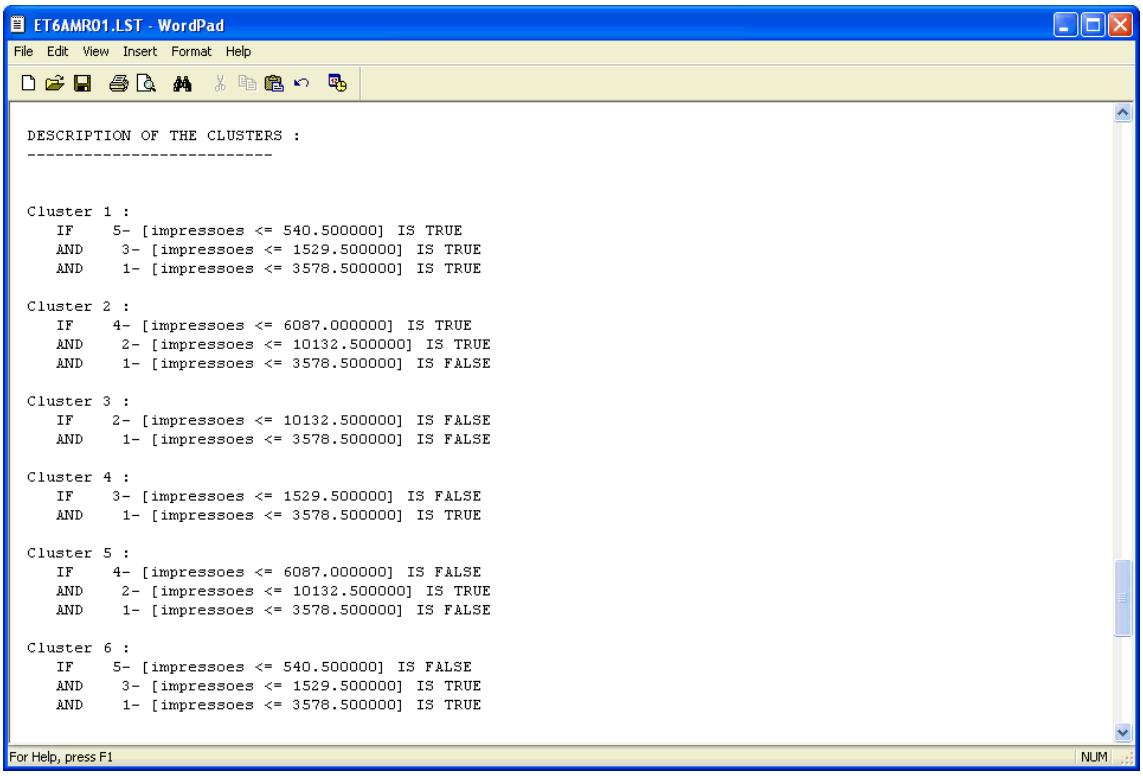


Figura 23 – *Log* dos dados da árvore com 6 *clusters* utilizando variável impressões (método DIV)

O método SCLUST permite que sejam gerados *clusters* misturando variáveis do tipo Modal e Interval na mesma análise. Este método diferentemente do DIV não efetua o corte baseado em

somente uma variável e efetua *clusters* baseado nos objetos simbólicos, o SCLUST é baseado nas informações de todas variáveis e efetua os *clusters* das mesmas. Pode-se verificar na figura 24, que foi utilizada para o *Data Mining* a distância Euclidiana e solicitados 6 *clusters*, fazendo uma análise com 20 interações, para todas análises efetuadas nesta dissertação com o método SCLUST é utilizada a distância Euclidiana.

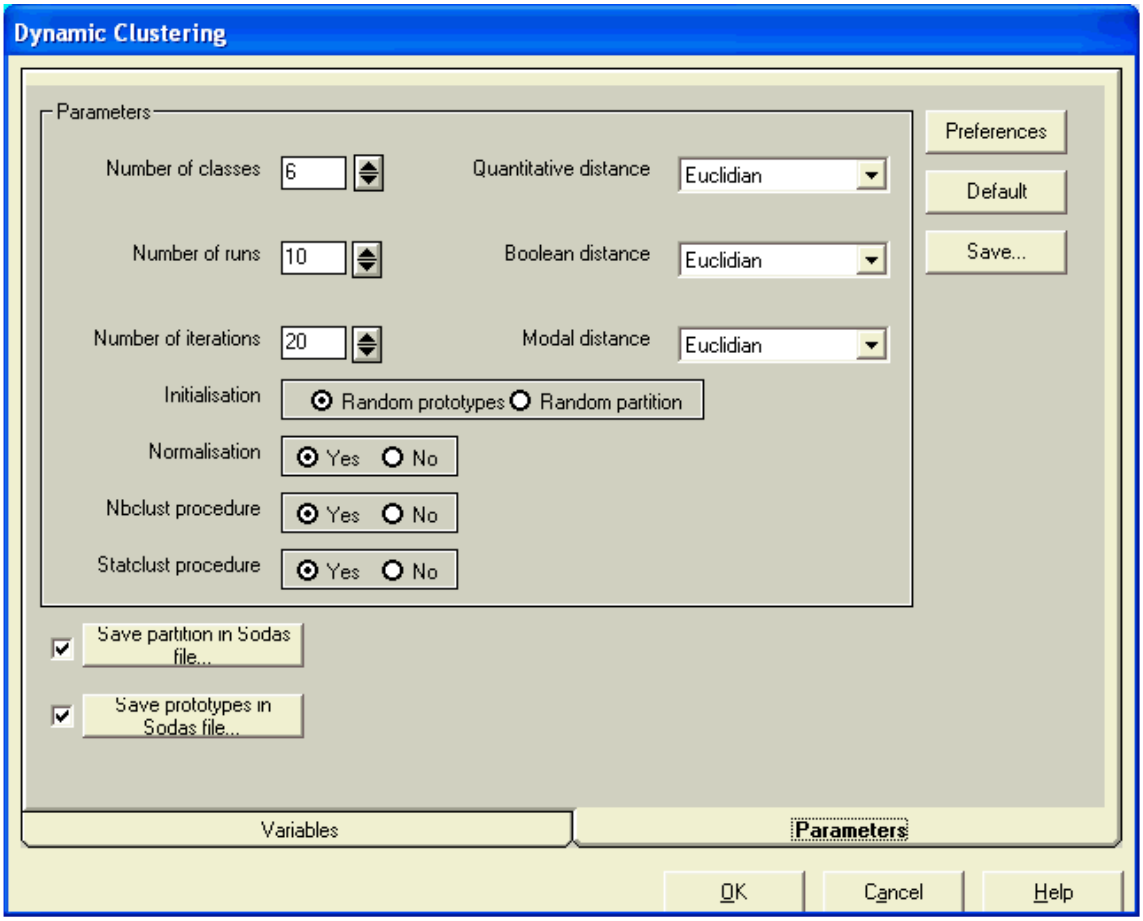


Figura 24 – Método S CLUST Selecionando Variáveis

Verificando o resultado na figura 25, pode-se ver claramente os *prototypes* formados, através de um gráfico "biplot" utilizando as variáveis "total_logins" e "impressões". Pode-se verificar que o *Prototype* 4/6 e *Prototype* 6/6 são os menos lucrativos para empresa, pois são os que geram menos resultados tendo em vista que estes usuários efetuam poucos *logins* no sistema, ou seja, acessam com menos freqüência o *site* e também geram uma quantidade muito pequena de impressões.

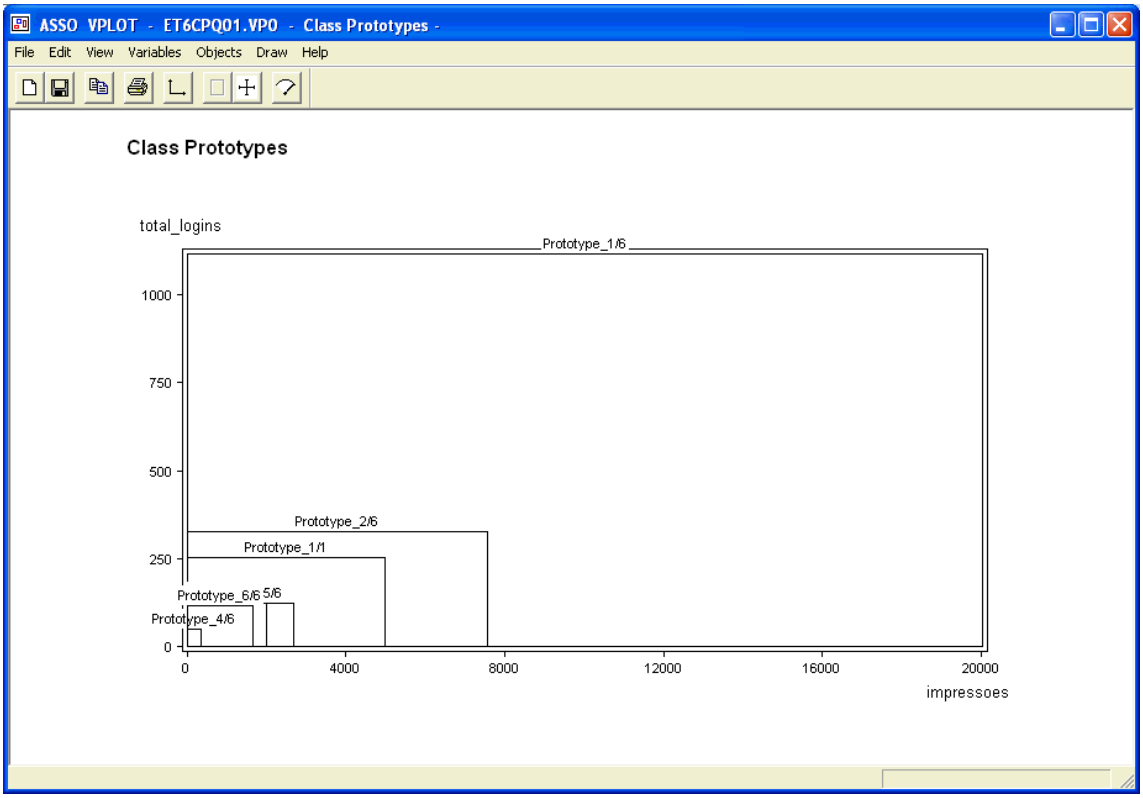
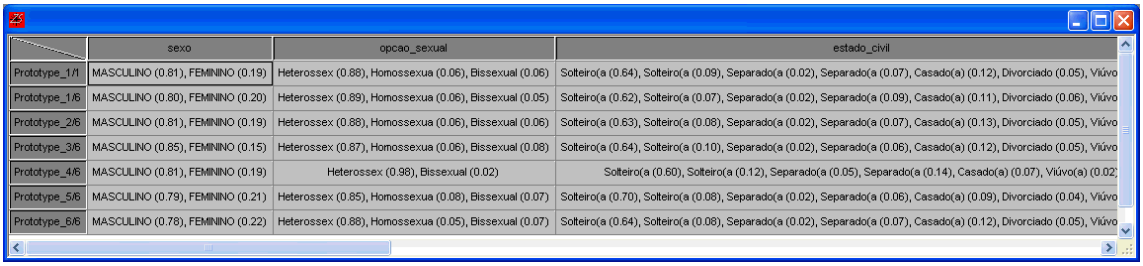


Figura 25 – C*lusters* baseados nas variáveis "Impressões" e "Total de Mensagens Enviadas". (método S CLUST)

O SCLUTS permite que seus resultados sejam importados para o método View, onde se pode identificar melhor como é composto cada *prototype*. Também deve ser bem avaliado os *prototypes* 1/6 e 2/6, pois são os que possuem os melhores resultados. (Fig. 26)



	sexo	opcao_sexual	estado_civil
Prototype_1/1	MASCULINO (0.81), FEMININO (0.19)	Heterossex (0.88), Homossexua (0.06), Bissexual (0.06)	Solteiro(a (0.64), Solteiro(a (0.09), Separado(a (0.02), Separado(a (0.07), Casado(a) (0.12), Divorciado (0.05), Viúvo
Prototype_1/6	MASCULINO (0.80), FEMININO (0.20)	Heterossex (0.89), Homossexua (0.06), Bissexual (0.05)	Solteiro(a (0.62), Solteiro(a (0.07), Separado(a (0.02), Separado(a (0.09), Casado(a) (0.11), Divorciado (0.06), Viúvo
Prototype_2/6	MASCULINO (0.81), FEMININO (0.19)	Heterossex (0.88), Homossexua (0.06), Bissexual (0.06)	Solteiro(a (0.63), Solteiro(a (0.08), Separado(a (0.02), Separado(a (0.07), Casado(a) (0.13), Divorciado (0.05), Viúvo
Prototype_3/6	MASCULINO (0.85), FEMININO (0.15)	Heterossex (0.87), Homossexua (0.06), Bissexual (0.08)	Solteiro(a (0.64), Solteiro(a (0.10), Separado(a (0.02), Separado(a (0.06), Casado(a) (0.12), Divorciado (0.05), Viúvo
Prototype_4/6	MASCULINO (0.81), FEMININO (0.19)	Heterossex (0.98), Bissexual (0.02)	Solteiro(a (0.60), Solteiro(a (0.12), Separado(a (0.05), Separado(a (0.14), Casado(a) (0.07), Viúvo(a) (0.02
Prototype_5/6	MASCULINO (0.79), FEMININO (0.21)	Heterossex (0.85), Homossexua (0.08), Bissexual (0.07)	Solteiro(a (0.70), Solteiro(a (0.08), Separado(a (0.02), Separado(a (0.06), Casado(a) (0.09), Divorciado (0.04), Viúvo
Prototype_6/6	MASCULINO (0.78), FEMININO (0.22)	Heterossex (0.88), Homossexua (0.05), Bissexual (0.07)	Solteiro(a (0.64), Solteiro(a (0.08), Separado(a (0.02), Separado(a (0.07), Casado(a) (0.12), Divorciado (0.05), Viúvo

Figura 26 – *clusters* na ferramenta View. (método SCLUST)

Para facilitar a análise, é possível extrair um gráfico em 3d do resultado e assim visualizar graficamente todos os *prototypes.* (Fig. 27).

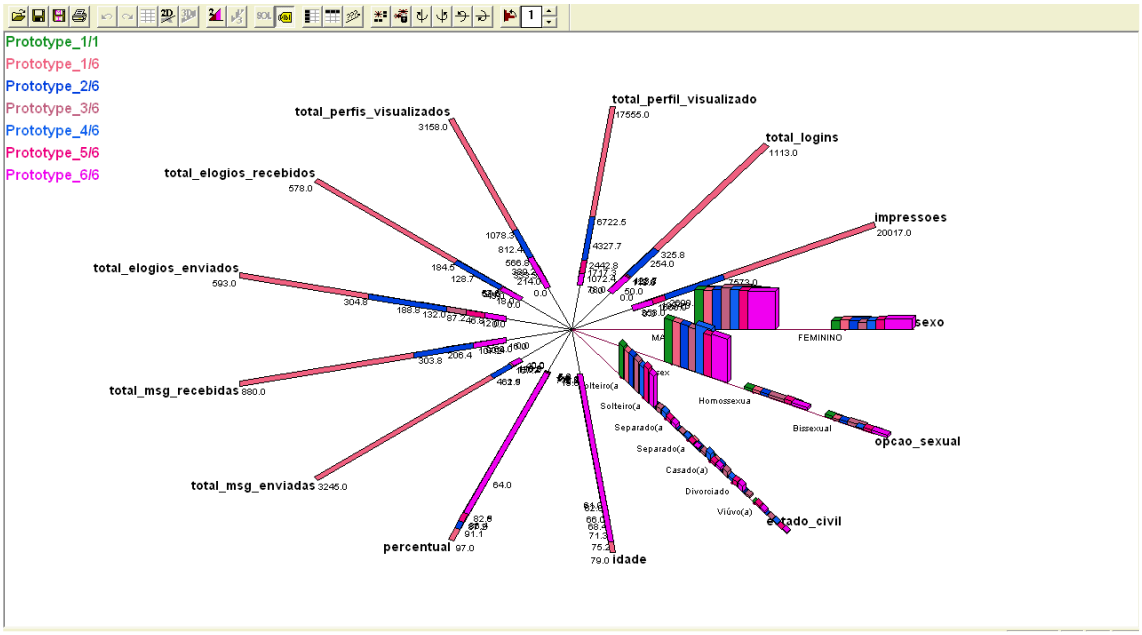


Figura 27 – Método S CLUST Selecionando Variáveis

Verifica-se que os *prototypes* 1/6 e 2/6, que são os de melhores resultados, possuem algumas características em comum. São grupos que possuem os maiores limites de idade, ou seja, possui um público não tão jovem e também são *clusters* de usuários que costumam ter um percentual maior de perfil preenchido o que faz com que os mesmos sejam mais atrativos, e que faz com que recebam mais elogios e mensagens de outros usuários. Estes dois *clusters* ótimos possuem a variável “sexo” e “opção sexual” semelhantes à média geral do site, fazendo com que estas duas variáveis não gerem muito impacto neste momento.

O próximo método de *clustering* que foi utilizado para o *Data Mining* do AondeNamoro.com é o Syksom, que possui em seu algoritmo características da construção de mapas de Kohonen.

O método Syksom foi utilizado na análise objetivando a formação de 6 *clusters* (2 linhas e 3 colunas na rede de Kohonen), tendo como base para os objetos simbólicos a variável “estado”, conforme pode ser observado na figura 28.

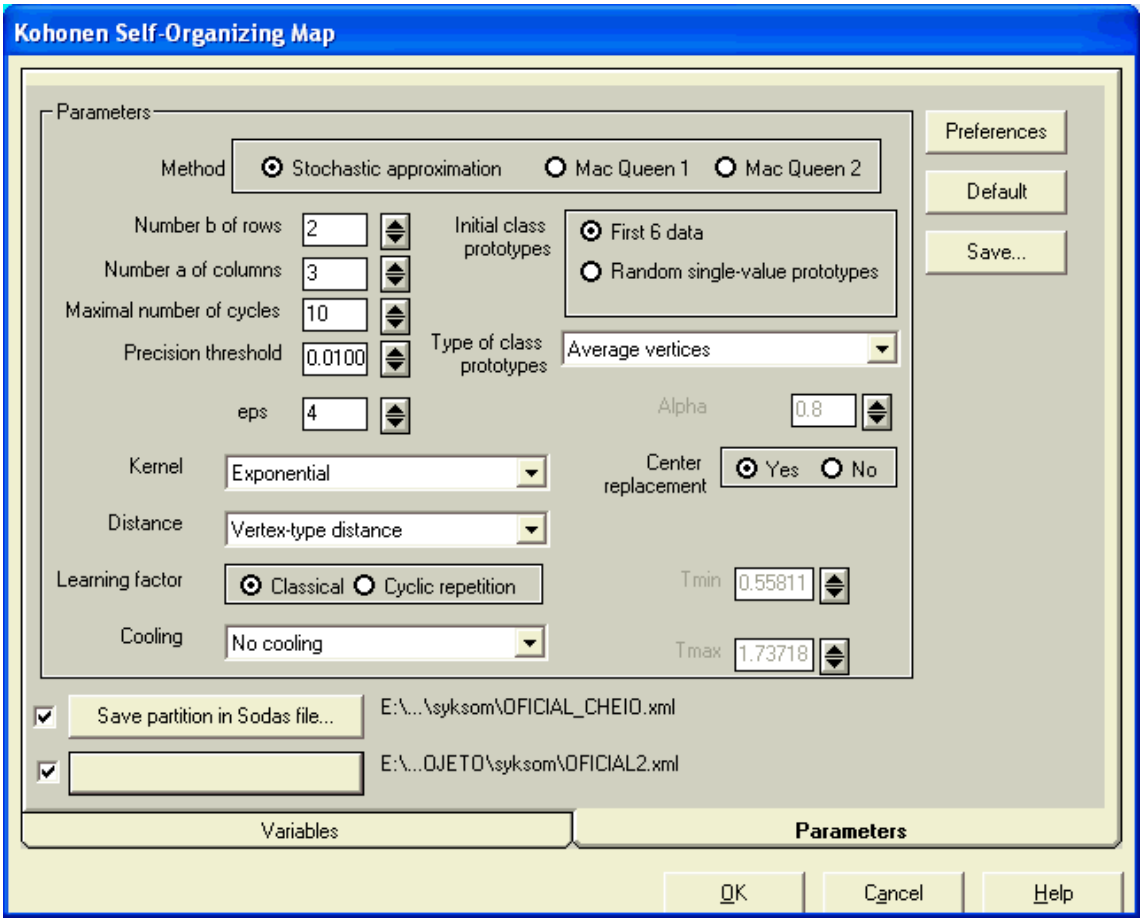


Figura 28 – Método Syksom – *Parameters*

O método Syksom, também trabalha somente com as variáveis do tipo Interval e utiliza todas elas para efetuar seus *clusters.* Tendo em vista os 6 *clusters* configurados, foi gerado um gráfico "biplot" onde se selecionou para análise inicialmente as variáveis "total_perfis_visualizados" e a variável "total_logins" (Fig. 29). Neste gráfico, são expostas algumas particularidades, onde se pode identificar que o *cluster* 2x2, apesar de realizar poucos *logins* no site do AondeNamoro.com, visualiza uma grande quantidade de perfis o que faz com que ele seja um usuário atrativo para empresa. O ideal é que o AondeNamoro.com incentive mais estes usuários a acessarem com mais freqüência o *site*, pois quando o fazem costuma permanecer um bom tempo no mesmo visualizando perfis e conseqüentemente gerando impressões e receita para empresa, existindo assim um grande potencial de

crescimento através da melhor exploração dos usuários do *cluster* 2x2. Facilmente também se identifica que o *cluster* mais atrativo para empresa é o 1x1, que efetua diversos *logins* no site e também visualiza uma grande quantidade de perfis. Já o *cluster* 1x2 aparentemente não é atrativo, tendo em vista que acessa poucas vezes o AondeNamoro.com e também visualiza uma pequena quantidade de perfis.

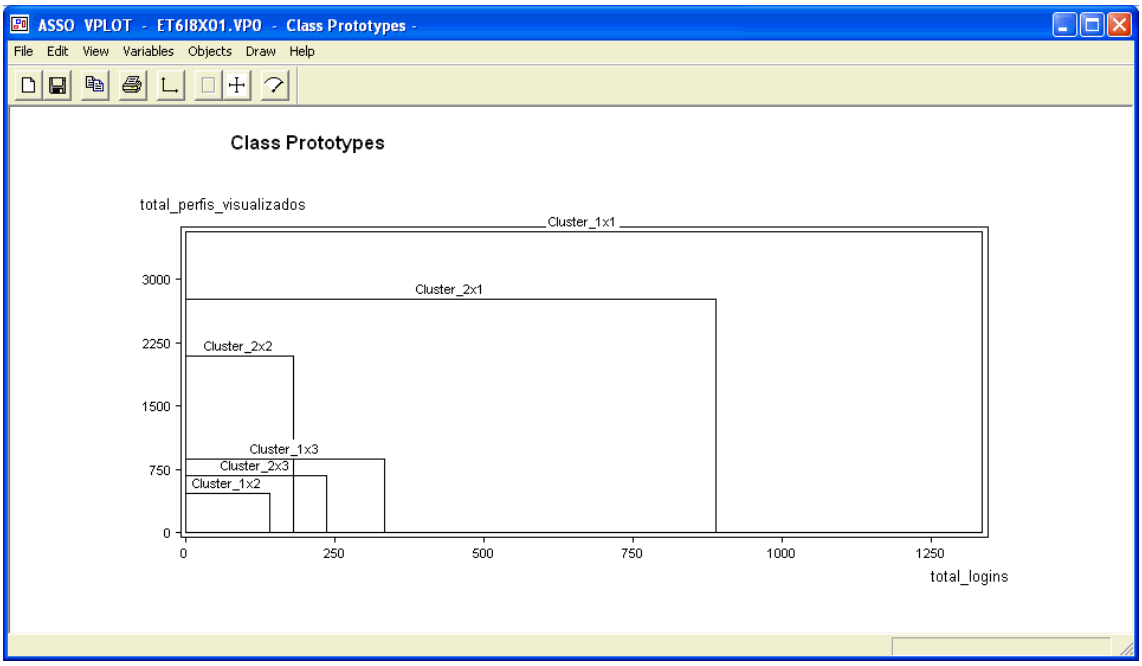


Figura 29 – Método Syksom – total_perfis_visualizados X total_logins

Através da análise da figura 30, que avalia a quantidade de impressões geradas e a quantidade de mensagens recebidas pelos *clusters*, pode-se perceber uma pequena alteração na importância dos *clusters*, o 1x1 que se torna aparentemente um pouco menos atrativo e o *cluster* 2x1 passa a ser o *cluster* mais atrativo, tendo em vista que seus usuários geram uma grande quantidade de impressões e também recebem muitas mensagens. Efetuando uma comparação do desempenho do 2x1 na figura 29 e 30, percebe-se que este *cluster* gera um bom resultado para empresa. Entretanto, o mesmo poderia ser maior, pois deveria ser

incentivado que acessasse com mais freqüência o AondeNamoro.com, pois o mesmo recebe uma grande quantidade de mensagens, porém a quantidade de *logins* efetuados não é tão grande como esperado. Verifica-se também, novamente, a importância do *cluster* 2x2, tendo em vista que geram uma grande quantidade de impressões e, proporcionalmente aos outros *clusters*, ele não recebe uma quantidade muito grande de mensagens de outros usuários.

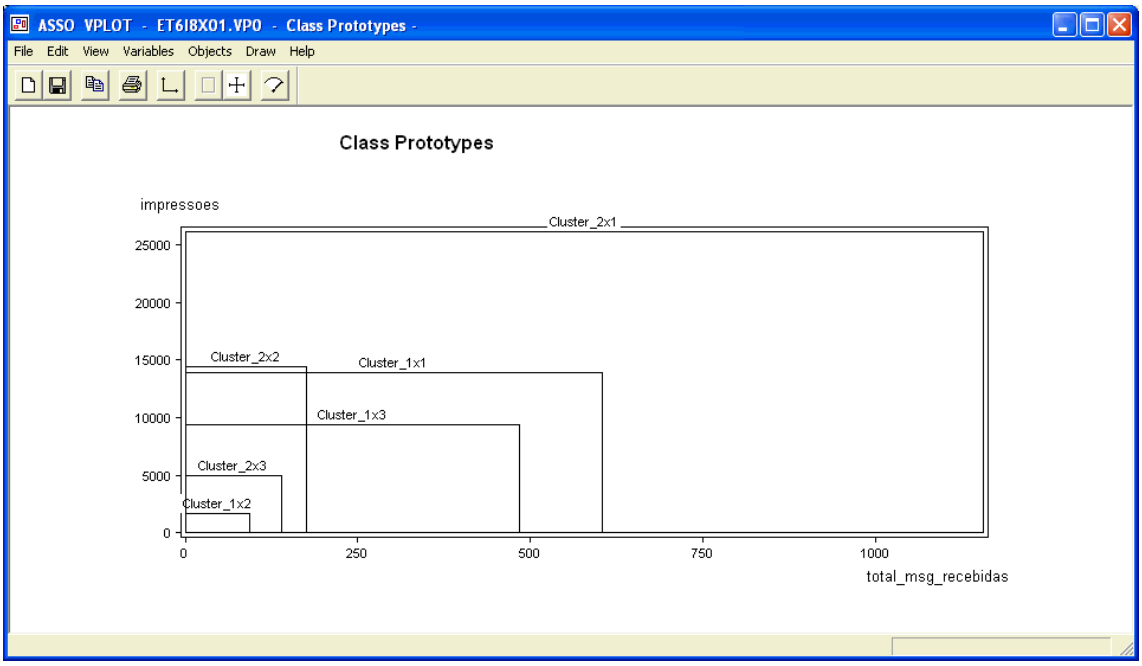


Figura 30 – Método Syksom – impressões x total_msg_recebidas

Analisando a figura 31, gráfico "biplot" da comparação do total de mensagens recebidas com a quantidade de vezes que o perfil do usuário foi visualizado, verifica-se que os usuários cadastrados no AondeNamoro.com possuem grande interesse em ter um relacionamento com os usuários do *cluster* 1x3, tendo em vista que, apesar do perfil dos mesmos serem pouco visualizados, eles geram uma grande quantidade de mensagens recebidas. Já o *cluster* 2x2, tem seus perfis acessados por muitos usuários do AondeNamoro.com. Entretanto, estes usuários por algum motivo não se interessam em enviar mensagens para os usuários deste

cluster. Seria interessante que o *site* efetuasse alguma ação junto a este *cluster* para ensinar os seus usuários a criarem perfis que estimulassem mais o envio de mensagens.

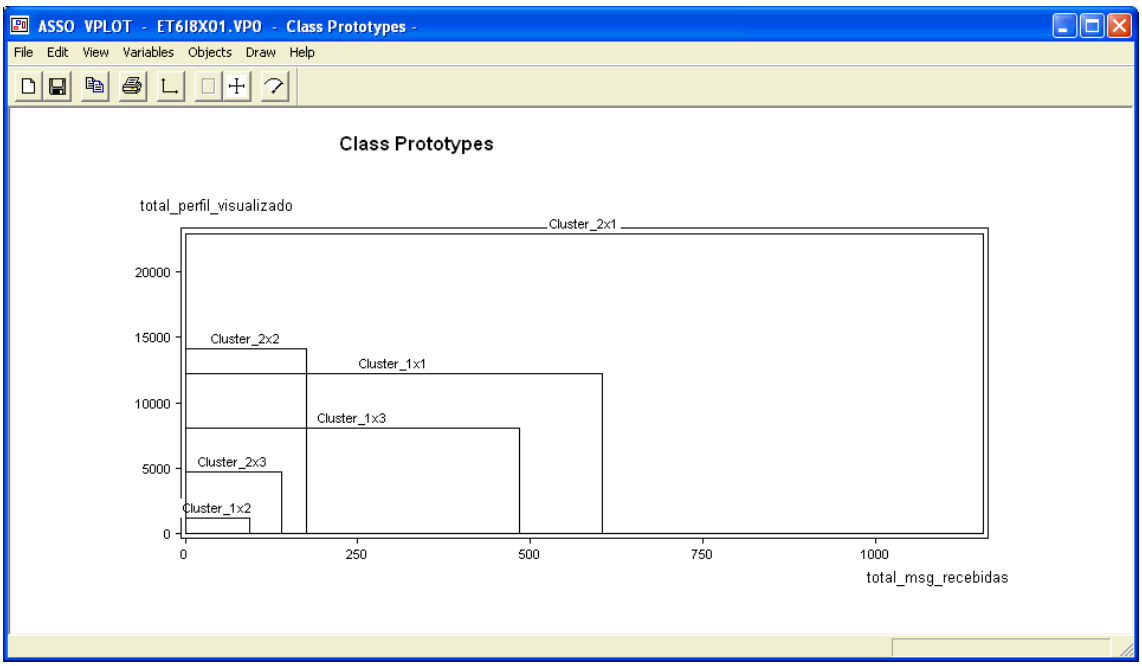


Figura 31 – Método Syksom – total_perfil_visualizado x total_msg_recebidas

Através da figura 32, pode-se extrair mais um conhecimento relevante do banco de dados do AondeNamoro.com, onde é efetuada uma análise comparando a quantidade de vezes que um determinado perfil é visualizado e o percentual de preenchimento que este perfil possui. Pode ser verificado que quanto mais preenchido um perfil estiver maior são as possibilidades dele ser visualizado por outros usuários. Observa-se, por exemplo, que *clusters* como o 1x2, que em praticamente todos os gráficos foi exposto como uma dos piores *clusters*, é composto pelos usuários que preenchem menos os seus perfis, o que claramente diminui a relevância do mesmo para nosso sistema e outros usuários.

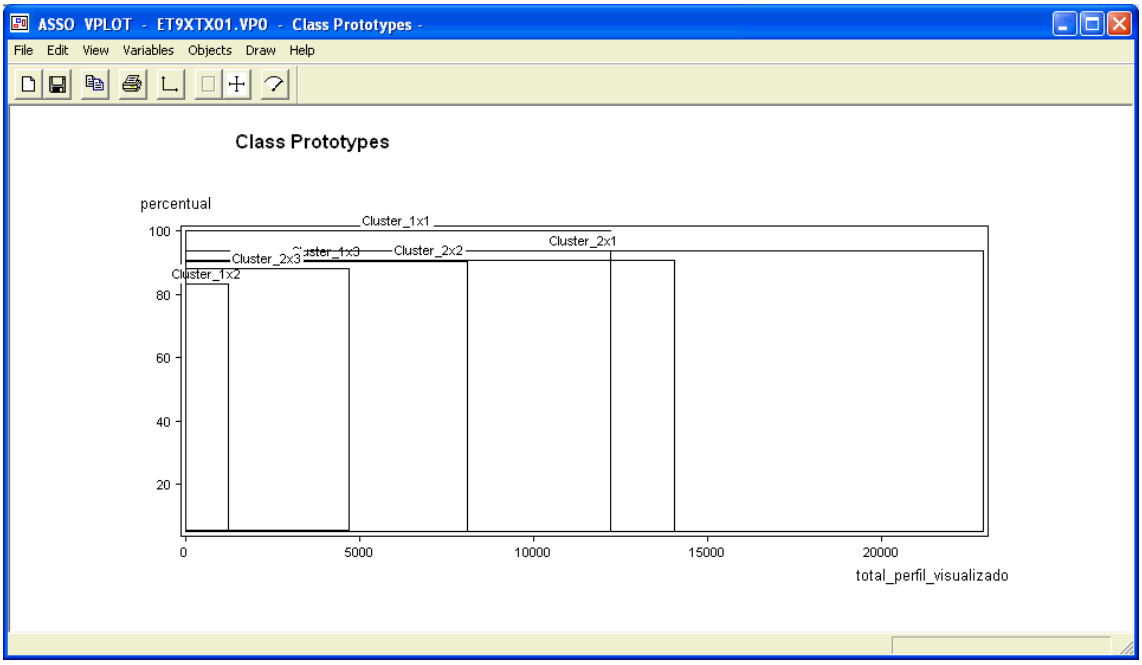


Figura 32 – Método Syksom – percentual X total_perfil_visualizado

Na figura 33, tem-se um resumo dos dados dos *clusters*, importado para ferramenta *View* do SODAS, sendo possível verificar as diferenças de cada *clusters*.

	idade	percentual	total_msg_enviadas	total_msg_recebidas	total_elogios_enviados	total_elogios_recebidos	total_perfis_visualizados	total_perfil_visualizado	total_logins	impressoes
Cluster_1x1	[18.00 : 72.00]	[5.00 : 100.00]	[0.00 : 1010.00]	[0.00 : 604.00]	[0.00 : 628.00]	[0.00 : 458.00]	[0.00 : 3552.00]	[0.00 : 12208.00]	[0.00 : 1336.00]	[0.00 : 13948.00]
Cluster_1x2	[18.19 : 66.25]	[5.63 : 83.38]	[0.00 : 185.88]	[0.00 : 92.63]	[0.00 : 112.50]	[0.00 : 53.13]	[0.00 : 468.00]	[0.00 : 1241.25]	[0.00 : 141.25]	[0.00 : 1700.75]
Cluster_1x3	[18.00 : 72.67]	[5.00 : 90.33]	[0.00 : 466.00]	[0.00 : 484.00]	[0.00 : 270.67]	[0.00 : 262.00]	[0.00 : 874.67]	[0.00 : 8092.00]	[0.00 : 334.67]	[0.00 : 9350.00]
Cluster_2x1	[18.00 : 86.00]	[5.00 : 94.00]	[0.00 : 5480.00]	[0.00 : 1156.00]	[0.00 : 558.00]	[0.00 : 698.00]	[0.00 : 2764.00]	[0.00 : 22902.00]	[0.00 : 890.00]	[0.00 : 26086.00]
Cluster_2x2	[18.00 : 82.00]	[5.00 : 91.00]	[0.00 : 218.00]	[0.00 : 176.00]	[0.00 : 418.00]	[0.00 : 170.00]	[0.00 : 2086.00]	[0.00 : 14066.00]	[0.00 : 182.00]	[0.00 : 14444.00]
Cluster_2x3	[18.20 : 65.60]	[5.40 : 88.40]	[0.00 : 283.60]	[0.00 : 140.40]	[0.00 : 176.40]	[0.00 : 102.80]	[0.00 : 684.40]	[0.00 : 4707.20]	[0.00 : 237.20]	[0.00 : 4990.40]

Figura 33 – Método Syksom – *Clusters* utilizados através da ferramenta *View*.

A figura 34 demonstra graficamente os dados de cada *cluster*, o que auxilia o AondeNamoro.com a conhecer seus nichos de usuários.

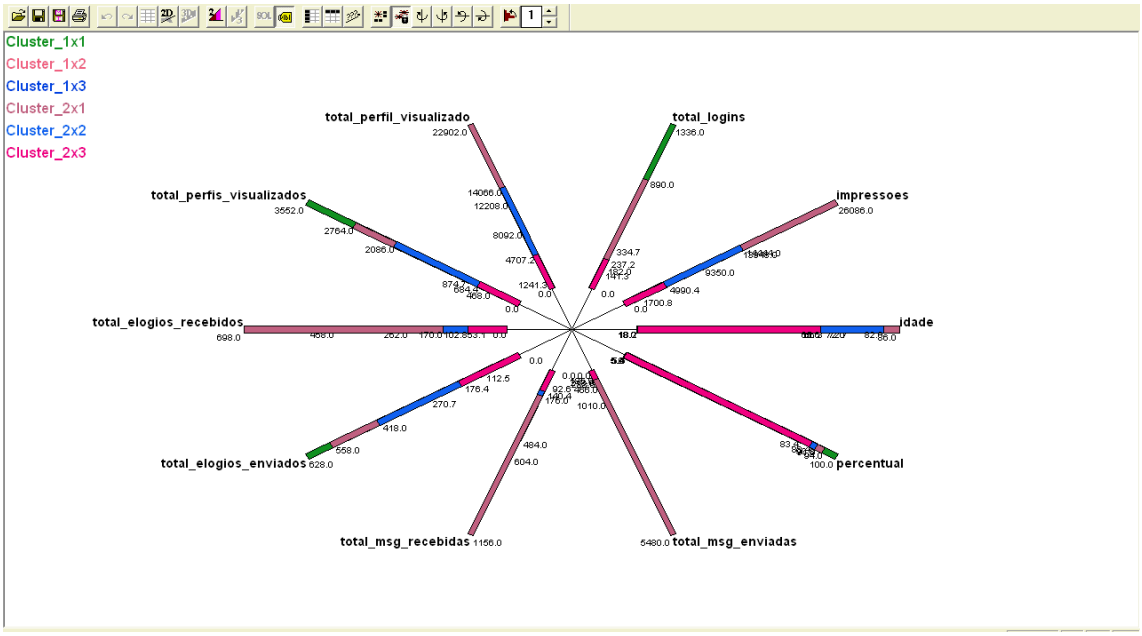


Figura 34 – Método Syksom – Visualização gráfica dos *Clusters*

Levando em consideração as análises efetuadas através do Método Syksom, o AondeNamoro.com deveria concentrar seus esforços nos *clusters* 2x1, 1x1 e 2x2, tendo em vista que os mesmos são os que demonstram os melhores resultados. Os *clusters* 1x2, 2x3 e 1x3 aparentemente são os de pior desempenho e menos atrativos para empresa, sendo aconselhável que se efetue um trabalho mais amplo para identificar o porque não estão sendo atingidas as necessidades destes nichos de usuários.

Através do método SCLASS, é possível identificar os estados que possuem mais semelhanças entre si, tendo em vista as variáveis Interval que são as únicas aceitas por este método. Este tipo de informação é muito importante para o AondeNamoro.com, principalmente para o desenvolvimento de sua campanha de divulgação, pois poderiam ser desenvolvidas peças específicas de publicidade (propaganda) para cada *cluster* de estados. Na figura 35 está exposta a árvore com os *clusters* dos estados.

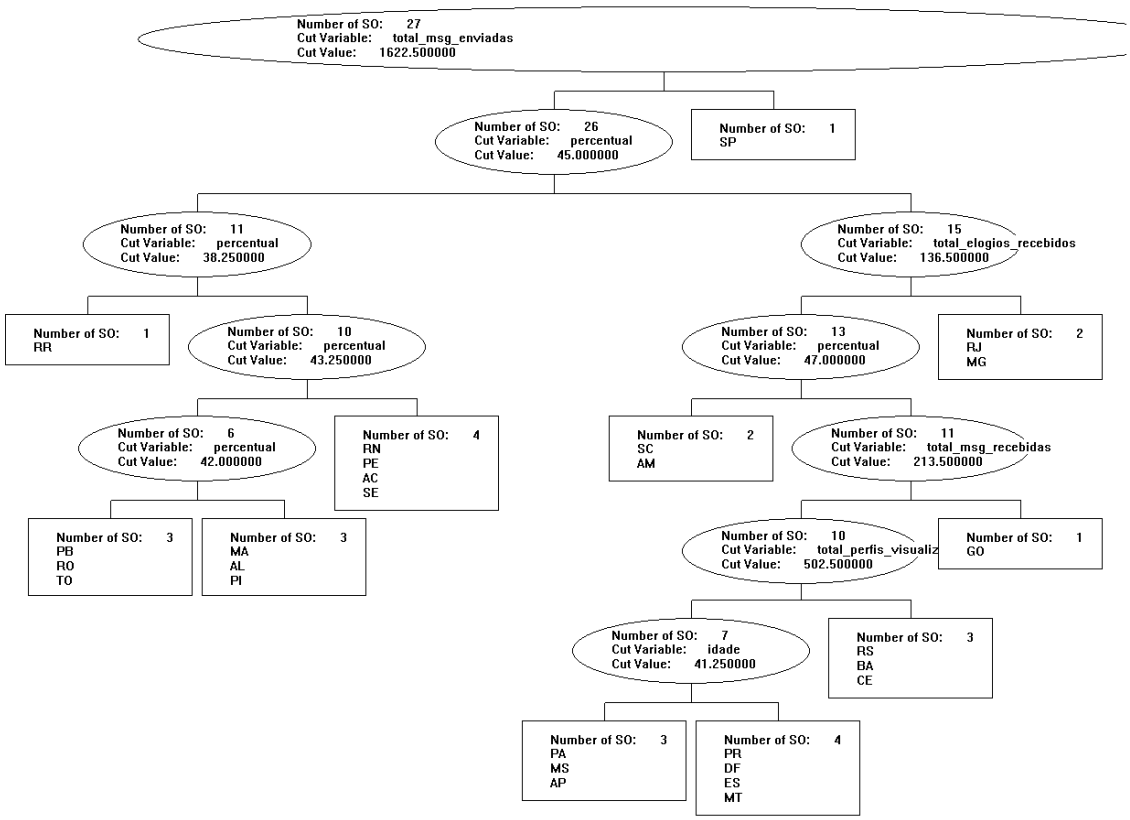


Figura 35 – Método SCLASS – Árvore de *clusters* de estados

Através da figura 35 é possível identificar 10 *clusters* de estados, que foram gerados através do algoritmo do SCLASS, que escolheu a variável ótima para o corte e o valor do corte para cada *cluster*. Sendo assim esta análise leva a acreditar que baseado nas variáveis do tipo Interval analisadas o AondeNamoro.com poderia desenvolver uma campanha de publicidade semelhante para os usuários de MG e RJ, tendo em vista que os mesmos possuem um comportamento semelhante dentro do site.

5 CONCLUSÃO

Este estudo teve como enfoque demonstrar a viabilidade da utilização do SODAS como software de *Data Mining*. Para alcançar este objetivo, foi desenvolvida a mineração da base de dados do site de relacionamento AondeNamoro.com e desenvolvido para o mesmo um modelo de *Data Mining.*

Dentro dos objetivos gerais, verifica-se que este estudo demonstrou como o SODAS é uma ferramenta de *Data Mining* eficiente e com a possibilidade de auxiliar a empresa AondeNamoro.com na tomada de decisões estratégicas.

Pode-se considerar que o estudo auxiliou a ampliar o campo do *Data Mining* para pequenas corporações, tendo em vista que demonstra como é possível a baixo custo, utilizar-se do SODAS para o conhecimento dos padrões de base de dados e a identificar padrões nos dados para auxiliarem em decisões estratégicas.

O estudo demonstra como através do *Data Mining* as empresas podem ampliar sua competividade, baseado nas informações estratégicas extraídas de suas próprias base de dados, transformando assim dado bruto em informação relevante.

O estudo foi eficaz no atingimento de seus objetivos específicos, onde foi desenvolvido um modelo de *Data Mining* para empresa estudada. Através da análise dos resultados, foram identificadas informações relevantes, que irão orientar a equipe do AondeNamoro.com durante suas mudanças estratégicas, do modelo de assinaturas para o modelo gratuito, e que possivelmente irão auxiliar a mesma a ampliar sua participação no mercado através do aumento de sua base de usuários, reconquistar usuários não mais ativos no site e ampliar a receita da empresa.

Outro objetivo específico muito importante atingido é o desenvolvimento de um tutorial "passo a passo" da utilização do SODAS (Apêndice A e B), que poderá auxiliar trabalhos futuros sobre o assunto.

A seguir serão identificadas as principais vantagens e desvantagens da utilização do *software* Francês SODAS, tão como as recomendações sugeridas a empresa estudada AondeNamoro.com e recomendações de trabalhos futuros.

5.1 VANTAGENS SODAS

Tendo em vista o estudo realizado com o SODAS, tanto para o desenvolvimento do passo a passo da utilização da ferramenta (Apêndices A e B) quanto no desenvolvimento do estudo de caso, é possível levantar as seguintes vantagens do SODAS (versão 2.5):

- Ferramenta livre;
- Interage com base de dados relacionais;
- Trata de forma competente os objetos simbólicos;
- Possui grande respaldo científico;
- Grande flexibilidade;

- Utilização de “Encadeamento” nos métodos;
- Possibilidade de trabalhar com grandes bases de dados;
- Disponível para *download* na Internet;
- Usabilidade intuitiva;

5.2 DESVANTAGENS SODAS

Foram identificadas as seguintes desvantagens na utilização do SODAS (versão 2.5):

- Necessidade de grande conhecimento em TI e também na linguagem SQL para interação com o Banco de Dados;
- Excesso de erros fatais no aplicativo. Por tratar-se de uma ferramenta ainda não amplamente testada, muitas vezes quando o SODAS localiza um erro e o programa é encerrado;
- Não possui uma uniformidade em todos sistemas. Em alguns computadores testados algumas funções retornavam erro em outros não;
- A opção “*Help*” do SODAS não abrange todas as opções dos *software*

5.3 RECOMENDAÇÕES A EMPRESA AONDENAMORO

- Criar outras variáveis para validar informações que podem ter sido informadas de forma equivocada propositalmente pelos usuários, como por exemplo, a variável "estado civil". Tendo em vista que informações incorretas no banco de dados podem enviesar as análises estratégicas realizadas;
- Considerando a correlação identificada na figura 19 entre, total de mensagens recebidas e enviadas pelos usuários, deve-se desenvolver peças de divulgação (propaganda) dentro do site, informando que usuários que enviam muitas mensagens

tendem a receber muitas, estimulando o envio de mensagens e conseqüentemente maior quantidade de impressões e receita para o *site*;

- Ampliar a base de usuários de sexo "FEMININO", tendo em vista que representam somente 20% da base de usuários e de acordo com o estudo são os usuários com maior potencial em gerar impressões e receita para empresa;
- Estimular os usuários a preencherem mais os seus perfis, pois de acordo com os dados obtidos, quanto maior o percentual de preenchimento de um perfil, mais o mesmo será atrativo para outros usuários, e consequentemente terá uma visualização maior o que gera mais impressões e receita par ao AondeNamoro.com (Figura 32);
- Desenvolvimento de campanhas de divulgação (propaganda) específicas para cada grupos de estado gerado através da análise utilizando o método SCLASS (Figura 35). Desta forma o AondeNamoro.com poderá economizar no custo de criação de publicidade e utilizar uma mesmo padrão de comunicação com públicos de iguais necessidades;
- Repetir o mesmo estudo com certa periodicidade para estar sempre em contato com as novas necessidades dos usuários e ter uma base histórica da evolução do perfil dos usuários.

5.4 TRABALHOS FUTUROS

- Monitorar o impacto das decisões do estudo realizado para o AondeNamoro.com;
- Monitorar o impacto da utilização do SODAS como ferramenta de apoio a decisão estratégica do AondeNamoro.com e no dia a dia da empresa;
- Realização de estudos demonstrando como o SODAS pode ser utilizado como uma ferramenta de *Business Intelligence* e assim auxiliar a gestão de pequenas e médias empresas;

- Desenvolvimento de tutoriais e estudos de casos que abordem os outros métodos do SODAS não abordados nesta dissertação;
- Comparação do SODAS com outras ferramentas de *Data Mining*;
- Comparação do custo benefício em se utilizar uma ferramenta de licença livre ou proprietária em *software*s de *Data Mining;*

REFERÊNCIAS BIBLIOGRÁFICAS

AGENTES no DI: projetos. Disponível em: <http://www.di.ufpe.br/~compint/projetos-agentes.html>. Data de Acesso: 5 de dezembro de 2006.

ANSELMO, Cézar Augusto de Freitas; CARVALHO, Francisco de Assis Tenório de; SOUZA, Renata Maria Cardoso Rodrigues. "Classificador Simbólico para Imagens SAR". In: SBSR, X, 2001 Foz do Iguaçu, Anais. Foz do Iguaçu, 2001 p. 21-26

ASSO Project: Analysis System of Symbolic Official data. Disponível em: < http://www.info.fundp.ac.be/asso/ >. Data de Acesso: 15 de outubro de 2006

BERSON, Alex; SMITH, Stephen; THEARLING, Kurt., "Bulding Data Mining Applications for CRM", ed. Mc Graw-Hill, EUA, 1999 ISBN 0-07-134444-6

BEZERRA, Byron Leite Dantas; QUEIROZ, Sérgio Ricardo de Melo. "Uma Visão de Análise de Dados Simbólicos." Disponível em: < http://www.cin.ufpe.br/~compint/aulas-IAS/kdd-022/SymbolicDataAnalysis.ppt>. Data de Acesso: 15 de agosto de 2006

BEZERRA, Dantas Leite Byron. "Estudo de Algoritmos de Filtragem de Informação Baseados em Conteúdo." Disponível em: <http://www.cin.ufpe.br/~tg/2001-2/bldb.pdf>. Data de Acesso: 7 de dezembro de 2006.

CARVALHO, Francisco de Assis Tenório. "Projeto CLADIS Classificação e Análise de Dissimilaridades." 2001. Disponível em: <http://ftp.cnpq.br/pub/protem/workshop2001/inria/relatorios/CLADIS.doc> Data de Acesso: 5 de dezembro de 2006.

CARVALHO, F.A.T; LECHEVALLIER, Y.. SCLUT Help Guide: Dynamic Clustering.. Sodas Versão 2.5. SCLUSTHLP.PDF, 20 de maio de 2003

CARVALHO, Francisco de Assis Tenório de; SOUZA, Renata M. C. R. de; SILVA, Fabio C. D. "Classificação Não Supervisionada de Dados de Tipo Intervalo baseada em Distâncias Não Quadráticas." Disponível em: <http://ftp.inf.pucpcaldas.br/CDs/SBC2003/pdf/arq0167.pdf>. Data de Acesso: 01 de dezembro de 2006.

CONSORTIUM. Asso Project. Disponível em: <http://www.info.fundp.ac.be/asso/consortium.htm>. Data de Acesso: 10 de agosto de 2006.

DIDAY, Edwin. “An Introduction to Symbolic Data Analysis and the SODAS Software.” Disponível em: <http://www.di.uniba.it/~malerba/activities/mod02/pdfs/diday.pdf>. Data de acesso: 10 de dezembro de 2006.

DISSEMINATION Activities. Asso Project. Disponível em: <http://www.info.fundp.ac.be/asso/dissemlink.htm >. Data de Acesso: 12 de agosto de 2006

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic; UTHURUSAMY, Ramasamy. “Advances in Knowledge Discovery and Data Mining” ed. MIT Press, EUA, 1996 ISBN 0-262-56097-6

FEATURES. Asso Project. Disponível em: <http://www.info.fundp.ac.be/asso/feature.htm>. Data de acesso: 10 de agosto de 2006.

HAIR, Joseph F.; ANDERSON, Rolph E.; TATHAM Ronald L.; BLACK, William C. “Multivariate Data Analysis with Readings”, edicao 4, ed. Prentice Hall International Editions, New Jersey, EUA 1995. ISBN 0-13-180969-5

HAN, Jiawei; KAMBER, Micheline. “Data Mining: Concepts and Techniques”, ed. Morgan KaufMann Publisher, 2001. Sao Francisco EUA. ISBN 4-55860-489-8

JOURNAL on Symbolic Data Analysis. Disponível em <http://www.jsda.unina2.it/>. Data de acesso: 10 de setembro de 2006

KAUFMAN, Leonard; ROUSSEEUW, Peter J.. “Finding Groups in Data: An Introduction to Cluster Analysis.” Ed. John Wiley & Sons, Inc. EUA. 1990. ISBN 0-471-87876-6

KENNEDY, Ruby L. et All. “Solving Data Mining Problems through Pattern Recognition”. ed. Prentice Hall, PTR, New Jersey, EUA. 1998. ISBN 0-13-095083-1

KNEIPP, Ricardo; ALBUQUERQUE, Rodney. “Criando Visões (Views) no Oracle.” SQL MAGAZINE. Disponível em: <http://www.sqlmagazine.com.br/Colunistas/RicardoRodney/02_ViewsNoOracle.asp> Data de Acesso: 31 de outubro de 2006

LAROUSSE. Dicionário da Língua Portuguesa. Ed. Ática. São Paulo, Brasil. 2001. ISBN 85-08-08077-8

LECHEVALLIER, Y.; CHAVENT, M. “DIV Help Guide: Divisive Classification.”. Sodas Versão 2.5. DIVHLP.PDF, 20 de maio de 2003

MAIORIA na web, mulheres procuram mais conteúdo multimídia, rádios e mapas. IdgNow. Disponível em: http://idgnow.uol.com.br/internet/2006/04/27/idgnoticia.2006-04-27.7382622157/IDGNoticia_view. Data de Acesso: 20 de novembro de 2006

MARTINS, Vidal. “Uma Visão Geral sobre ODBC.” Disponível em: <http://www.pr.gov.br/batebyte/edicoes/1996/bb53/odbc.htm>. Data de acesso: 12 de dezembro de 2006

MENESES, Esteban; Rodríguez-Rojas. “Using Symbolic Objects to Cluster Web Documents”. XV World Wide Web Conference. 2006. Disponível em:

<http://www2006.org/programme/files/pdf/p115.pdf>. Data de acesso: 01 de dezembro de 2006

METZ, Jean; MONARD, Maria Carolina. “Clustering hierárquico: uma metodologia para auxiliar na interpretação dos clusters.” In: XXV Congresso da Sociedade Brasileira de Computação, 2005. São Leopoldo. UNISINOS. 2005

MYSQL. SearchOpenSource.com. Disponível em: <http://searchopensource.techtarget.com/sDefinition/0,,sid39_gci516819,00.html>. Data de acesso: 12 de dezembro de 2006.

OBJECTIVES. Asso. Disponível em <http://www.info.fundp.ac.be/asso/objective.htm> . Data de acesso: 01 de agosto de 2006

PARTICIPANTS Sodas. Ceremade. Disponível em: < http://www.ceremade.dauphine.fr/%7Etouati/participantsSODAS.htm>. Data de aceso: 15 de agosto de 2006.

PRESS, S. James. “Applied Multivariate Analysis.” Ed. Holt, Rinehart and Winston, Inc. EUA. 1972. ISBN 0-03-082939-9

PUBLICATIONS Sodas. Disponível em: <http://www.ceremade.dauphine.fr/%7Etouati/biblio.htm>. Data de Acesso: 12 de agosto de 2006.

PYLE, Dorian, “Data Preparation for Data Mining”, ed. Morgan Kaufmann Publishers, Sao Francisco CA EUA, 1999 ISBN 1-55860-529-0

RIBEIRO, Luiz Alberto Pereira Afonso. “ Uso de Data Mining no apoio à gestão de qualidade e risco nas operações de captura em um banco apoiado no modelo de correspondentes bancários - o caso do Lemon Bank.**”**. 2005. Dissertação (Mestrado em Administração) - IBMEC, Rio de Janeiro, 2003

RUD, Olivia Parr, “Data Mining Cookbook Modeling Data for Marketing, Risk and Customer Relationship Management”, Wiley Computer Publishing, EUA ,2001. ISBN 0-471-38564-6

SCIENTIFC Background. Asso. Disponível em: <http://www.info.fundp.ac.be/asso/scientific.htm>. Data de acesso: 05 de agosto de 2006.

SCLASS Help Guide: Unsupervised Classification Tree. Sodas Versão 2.5. SCLASSHLP.PDF, 2 de outubro de 2003

SILVA, Marcelino Pereira dos Santos. “Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka.” SBC - Sociedade Brasileira de Computação. Disponível em: <http://www.sbc.org.br/bibliotecadigital/download.php?paper=35>. Data de Acesso: 12 de dezembro de 2006.

“SODAS: List of Treatments and Modules for Display in the Workbench.” Sodas Versão 2.5. SODASHLP.PDF, 04 de abril de 2004

SODAS-PAGEGARD. Dauphine Ceremade. Disponível em <http://www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm> . Data de acesso: 01 de agosto de 2006.

SYKSOM Help Guide: Kohonen Self-Organising. Sodas Versão 2.5. SYKSOMHLP.PDF, 6 de junho de 2003

THURAISIGHAM, Bhavani. “Data Mining Technologies, Techniques, Tools, and Trends.” ed CRC PRESS LLC, Florida, EUA, 1998. ISBN 0-8493-1815-7

TOP *Sites* Portuguese. Alexa. Disponível em: < http://www.alexa.com/site/ds/top_*sites*?ts_mode=lang&lang=pt>. Data de acesso: 28 de novembro de 2006.

WEISS, Sholom M.; INDURKHYA, Nitin. “Predictive Data Mining: a pratical guide”. ed. Morgan Kaufmann Publishers, são Francisco, EUA. ISBN 1-55860-403-0. 1998

WHAT are outliers in the data. “Engineering Statistics HandBook”. Disponível em: < http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>. Data de acesso: 15 de outubro de 2006.

WORKPACKAGES. Asso Project. Disponível em: <http://www.info.fundp.ac.be/asso/wpackage.htm>. Data de Acesso: 10 de agosto de 2006.

APÊNDICE A

O SODAS possui uma ferramenta específica chamada DB2SO, que é instalada juntamente com o mesmo. Esta ferramenta é responsável por efetuar toda a importação de dados e também a manipulação destes dados, como por exemplo, para criação de taxonomias e regras.

O primeiros passo para importar dados é abrir a Ferramenta do DB2SO, que é acessível através do SODAS (Figura 36) no "*Sodas File -> Import -> Import with DB2SO*". Conforme pode-se verificar na figura 37.

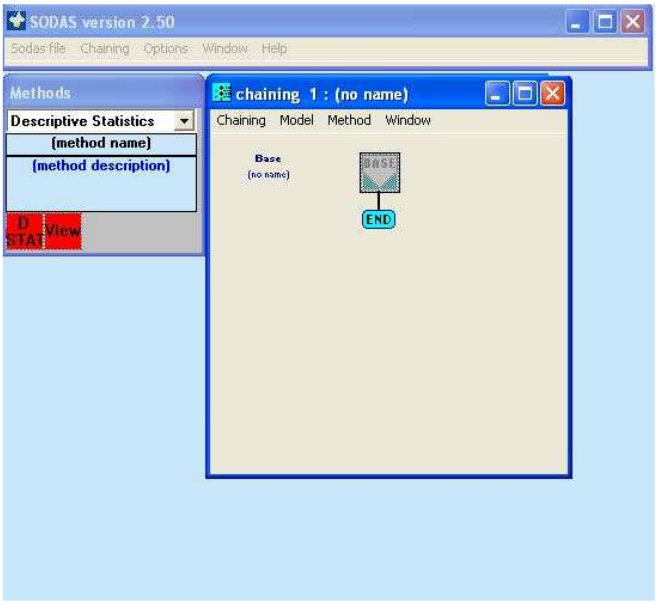


Figura 36 – Tela inicial do SODAS

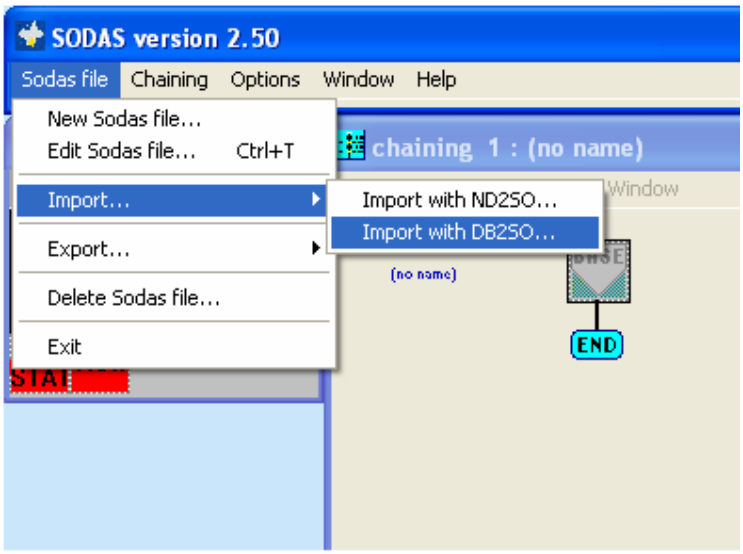


Figura 37 – Opção do Menu para abrir o aplicativo DB2SO

Após ter clicado em *"Sodas File -> Import -> Import with DB2SO"*, é aberto o DB2SO, conforme pode ser visualizado na figura 38.

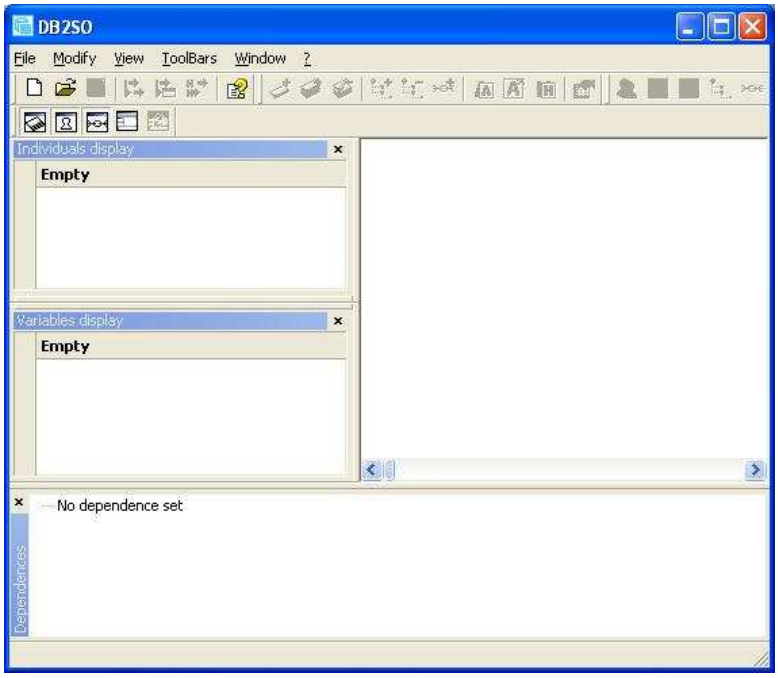


Figura 38 – Tela inicial do aplicativo DB2SO

Uma vez no DB2SO, é preciso se conectar no banco de dados para poder de fato iniciar o processo de importação dos dados para o SODAS. Para isto basta selecionar uma Fonte de Dados, clicando no menu "*File -> New*", sendo exposta a janela "*Select Data Source"*, onde se pode escolher o *Data Source* a ser utilizado (Fig. 39).

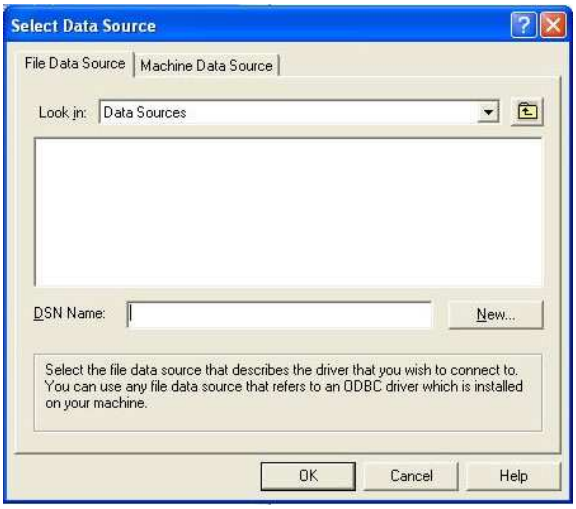


Figura 39 – Tela para escolha do Data Source

Conforme exposto no exemplo do estudo de caso do AondeNamoro.com é utilizada uma conexão com o Mysql através de ODBC. Sendo assim deve-se clicar na Aba "*Machine Data Source*" e selecionar o tipo de dado desejado, neste caso a opção é "namoro_trabalhado", que já se encontra configurado (Figura 40). Entretanto, também são permitidas importações de dados de Access (*.mdb), Excel (*.xls), entre outros.

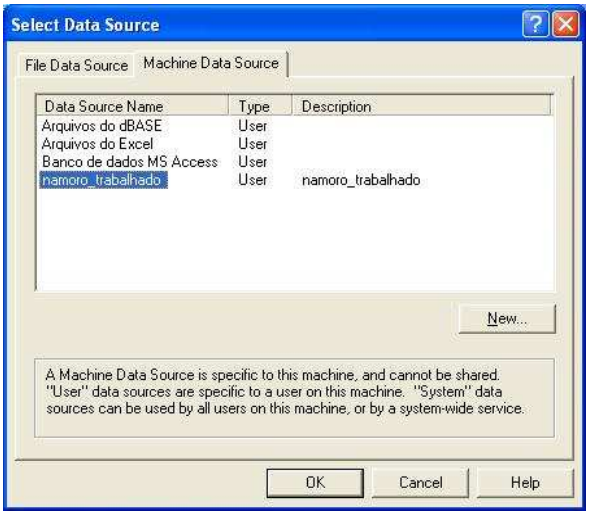


Figura 40 – Tela para escolha de *Data Source* configurado no ODBC

Pode-se verificar na figura 41**,** uma das etapas mais importantes da importação dos dados. Através da janela "*Extraction of Individuals*", onde é informado para o DB2SO, quais variáveis de quais tabelas se deseja utilizar no estudo. Nesta janela é exibida uma relação de todas as tabelas existentes e todas *queries* disponíveis para serem utilizadas na importação dos dados. Caso já tenha sido configurada uma *query* no caso do Access ou uma *View* utilizando o Mysql, a importação é efetuada de forma mais simples, onde se torna necessário somente selecionar as Tabelas ou *Queries* e clicar em "ok".

Por tratar-se de um *software* de *Data Mining*, o SODAS, possui nesta fase no campo "*Settings*" (Fig. 41) a opção de dividir sua base de dados por amostras de n elementos, podendo o usuário trabalhar livremente com estes valores, visando assim reduzir a massa de dados utilizada, tornando o processo de manipulação dos dados mais simples e rápido.

Figura 41 – Tela para seleção de tabelas, *queries* ou *views* para utilizadas pelo DB2SO

Na ferramenta de importação do SODAS, pode-se visualizar informações das tabelas, visando assim facilitar a utilização da ferramenta para consultas personalizadas. Para visualizar informações sobre suas tabelas ou *queries*, o usuário, deve selecioná-las e clicar no botão "*VIEW*" presente na figura 41 e assim são expostas as informações da tabela ou *query*, conforme figura 42.

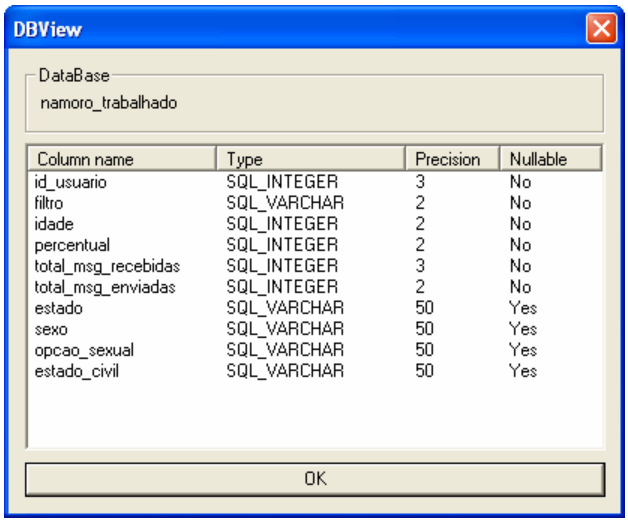


Figura 42 – Descrição gerada pelo DB2SO de uma tabela, *query* ou *view*

O DB2SO, por tratar-se de uma ferramenta que se conecta diretamente com diversos tipos de SGDB, também oferece uma excelente solução, para usuários experientes e que desejam efetuar uma importação de dados mais personalizada. Oferecendo a possibilidade de buscar dados de várias tabelas e ou *queries* (ou *Views*) ou então somente alguns campos de uma tabela e não a mesma inteira, isto através da opção "*Write Query*" (Figura 43), onde o usuário do *software* pode escrever precisamente o que deseja através de linguagem SQL. Este recurso é muito interessante, pois oferece liberdade ao Usuário do sistema para desenvolver as *queries* desejadas no momento em que esta trabalhando no DB2SO, não sendo necessário que ele crie uma *View* ou consulta no Access, sempre que desejar selecionar uma determinada massa de dados diferente da configurada.

Figura 43 – Tela para digitação de *query* no DB2SO

Após escrever seu comando SQL ou selecionar a tabela, *query* ou *View* desejada e clicar no botão de "OK", o DB2SO busca todos os dados informados e os organiza para o padrão do SODAS. Através da figura 44 pode ser verificado o *output* gerado pelo DB2SO, da

importação dos dados, mostrando informações básicas relevantes, como, por exemplo: tempo demorado para importação, quantidade de registros analisados e quantidade de variáveis qualitativas e quantitativas. Pode ser analisada também na área esquerda do DB2SO a relação das variáveis importadas e também das variáveis de classificação dos dados e na parte inferior do DB2SO são exibidas as regras de dependências, que serão abordadas mais a frente.

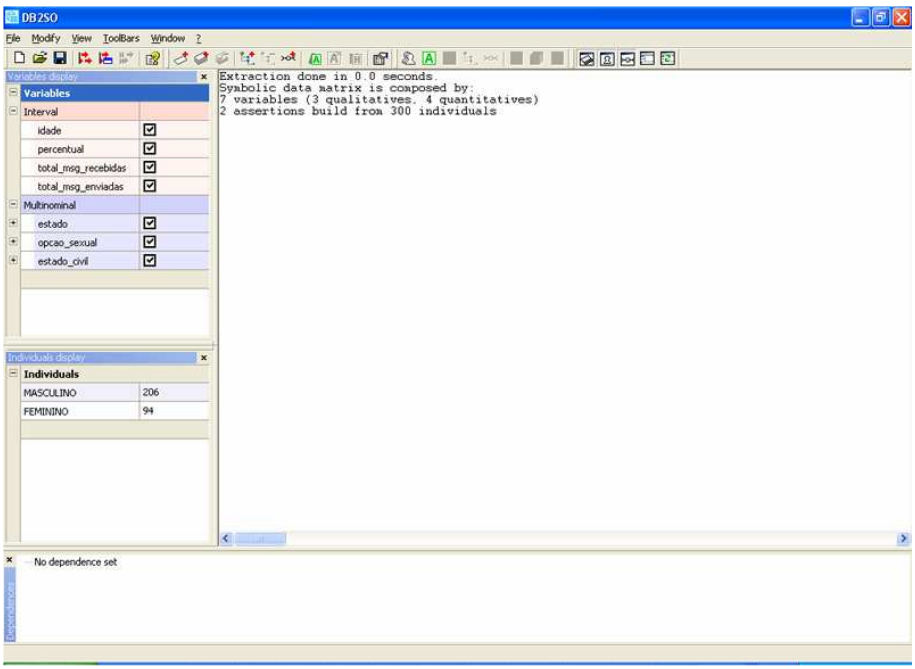


Figura 44 – *Output* gerado pela importação realizada pelo DB2SO

A importação dos dados para o SODAS ocorre até a etapa demonstrada na Figura 45. Entretanto, conforme abordado anteriormente, a ferramenta DB2SO possui algumas funcionalidades importantes para manipulação dos dados após a importação, sendo as mais importantes: Criação de Taxonomia e Criação de Regras de Dependência.

Segundo o dicionário Larousse (2001, pág. 951), taxonomia é a “teoria das classificações”.

Para criar uma Taxonomia, o ideal é que o usuário crie uma tabela no Banco de Dados com as taxonomias que deseja utilizar em seu modelo. Após criar esta tabela, o Usuário deve acessar no menu do DB2SO a opção "*Modify -> Create a Taxonomy*", onde é aberta a Janela representada na figura 45, para que o usuário selecione a variável que terá uma taxonomia criada, no exemplo "estado", e selecionar a tabela ou *query* de taxonomias e clicar em "OK". Pode também ser desenvolvida uma *View* para cada taxonomia, como no estudo de caso.

Figura 45 – Criação de Taxonomia

Após criar suas taxonomias, o usuário poderá verificar quais foram as Taxonomias criadas em seu modelo, para isto deverá acessar no menu a opção "*View -> Taxonomies*", sendo assim exposta de forma bem prática todas as taxonomias até então criadas (Figura 46).

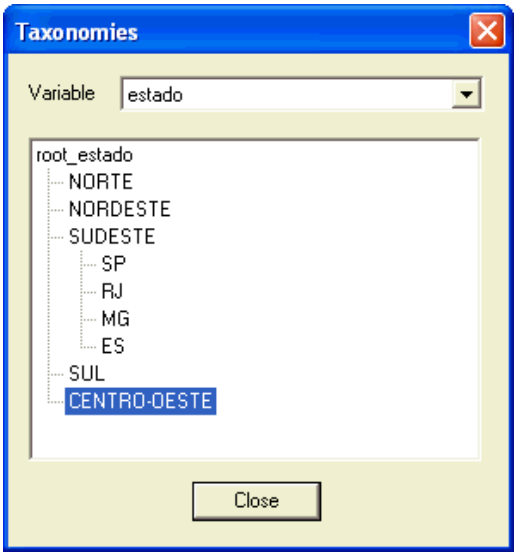


Figura 46 – Relação de Taxonomias criadas

Outra funcionalidade muito interessante disponível no DB2SO é a possibilidade de criação de dependência de variáveis. Esta funcionalidade deve ser utilizada, quando se deseja utilizar uma variável somente quando uma outra variável assumir um determinado valor. Para adicionar uma dependência, deve-se acessar no menu a opção *"Modify -> Add a dependence"*, sendo então aberta a janela representada aqui pela figura 47. Conforme se pode verificar, deve-se configurar três opções: *"Select mother variable" (*variável mãe), *"Select applicable set"* (valor da variável mãe que será utilizado) e *"Select daughter variables"* (variáveis filhas, que somente serão utilizadas no modelo quando a variável mãe possuir o valor selecionado).

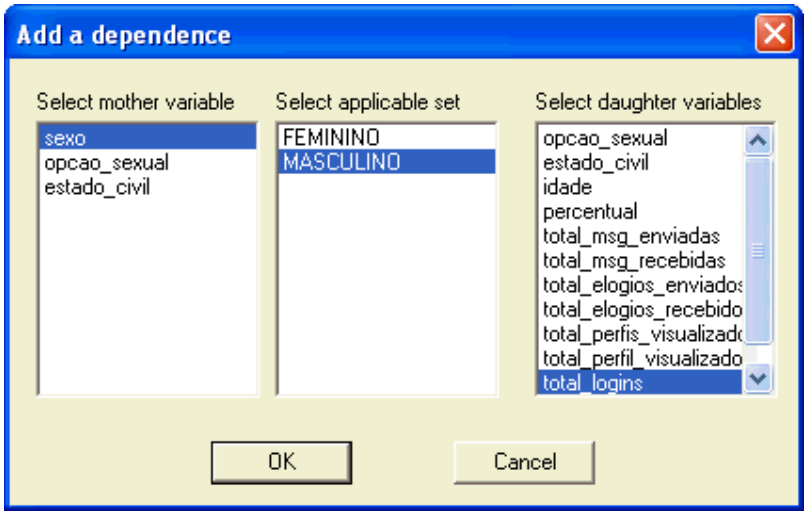


Figura 47 – Criação de Dependência

Após selecionar a configuração da dependência, o usuário deve clicar no botão de “ok” e assim é exposto na tela o *output* da dependência. Pode-se verificar no *output*, que foi adicionada 1 regra, onde somente é aplicada a variável (filha) “total_logins”, quando o valor da variável (mãe) “sexo” possuir o valor de “MASCULINO” e quando o valor de “sexo” for diferente os DB2SO substitui seu valor por “NA” . Caso o modelo possua diversas dependências, pode ser visualizado um consolidado das mesmas no formato de árvore logo abaixo do resultado em texto. (Figura 48)

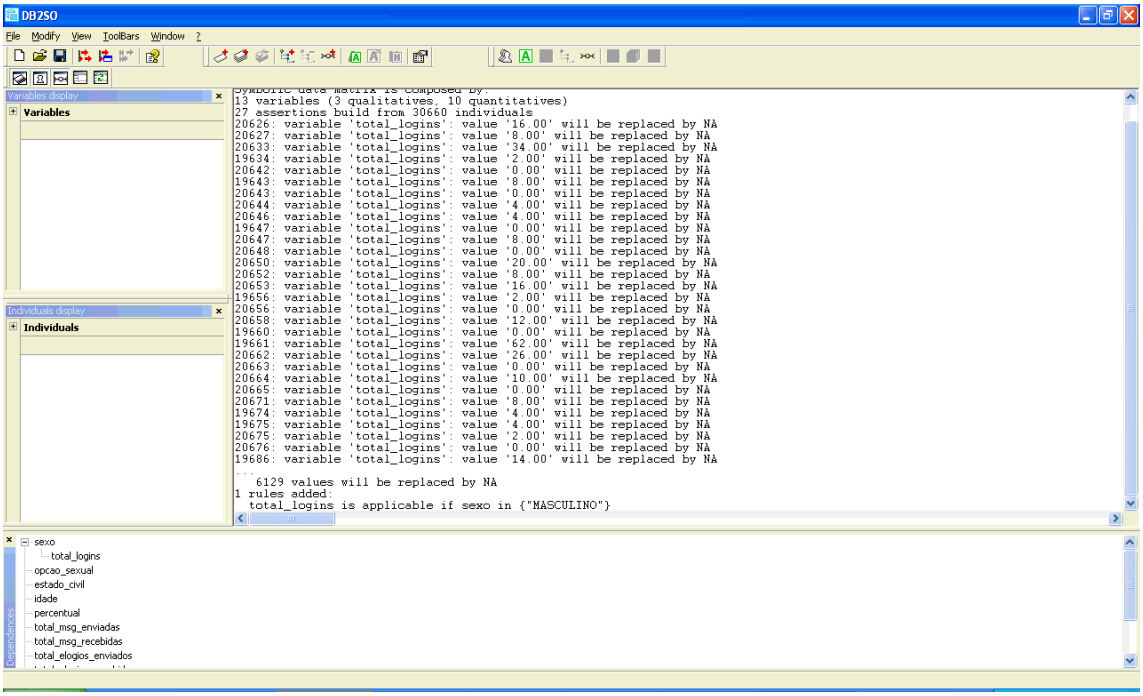


Figura 48 – Criação de Dependência

APÊNDICE B

ESTATÍSTICA DESCRITIVA

Neste tópico tem-se uma visão mais prática do funcionamento do SODAS para *data mining* , inicialmente através do passo a passo de como utilizar a funcionalidade de estatística descritiva. São abordados a seleção da database, que é necessária para qualquer análise, e também a utilização de qualquer módulo e também os módulos VIEW e DSTAT

VIEW

O módulo *View*, como seu próprio nome ressalta, tem como funcionalidade principal demonstrar através de diversos gráficos em 2d e 3d informações básicas sobre os dados simbólicos e as variáveis utilizadas no modelo.

Para utilizar o módulo *View* ou qualquer outro módulo, deve-se inicialmente "selecionar" um banco de dados, através de um arquivo "*.sds" (extensão utilizada pelo SODAS), que foi gerado durante a importação dos dados para o SODAS. Este arquivo "*.sds" pode ser visualizado através de qualquer editor de texto, como por exemplo o "Bloco de Notas", existindo no mesmo um resumo das informações coletadas, consolidadas através de matrizes para serem utilizadas pelo SODAS (Fig. 49). O SODAS também gera outros tipos de

arquivos, como por exemplo “*.gaj”, que possuem as informações completas para efetuar o *Data Mining*, entretanto eles somente podem ser interpretados pelo SODAS.

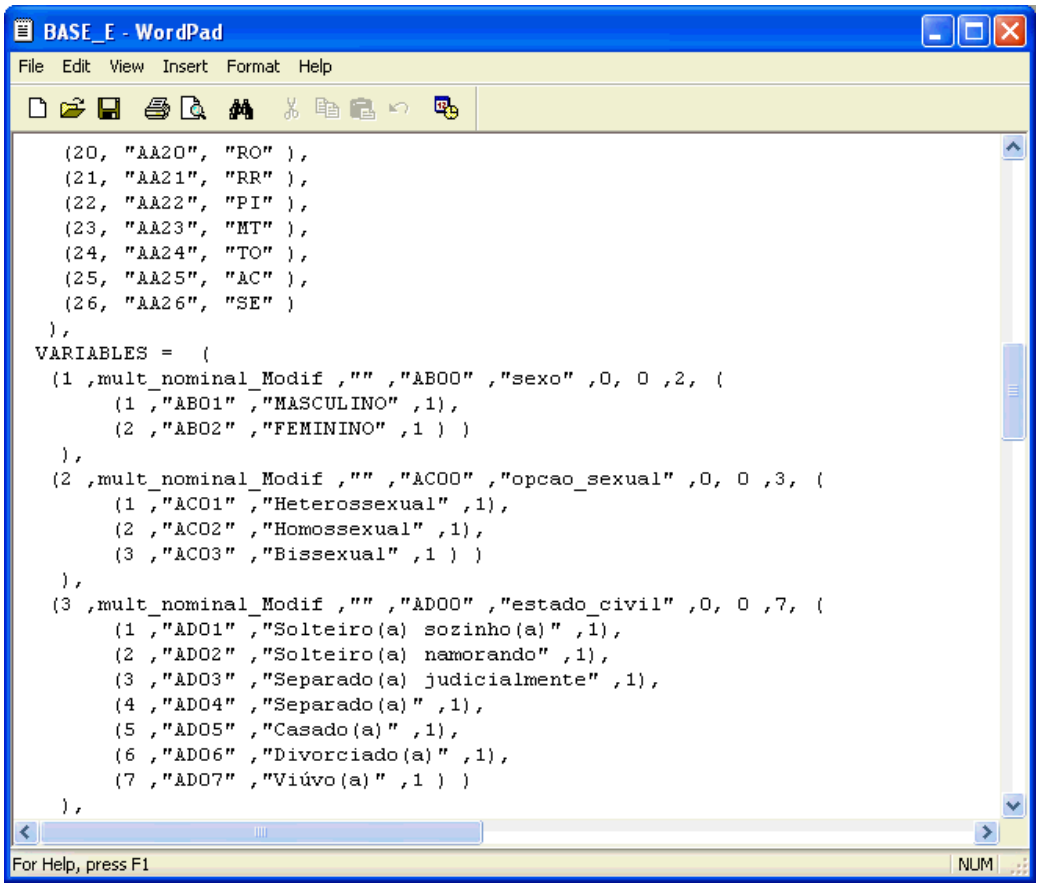


Figura 49 – Exemplo de arquivo “.sds”

A seleção do arquivo “*.sds” é efetuada através da opção “*Chaining -> Select Base*” na janela “*Chaning*” do SODAS (Fig. 50).

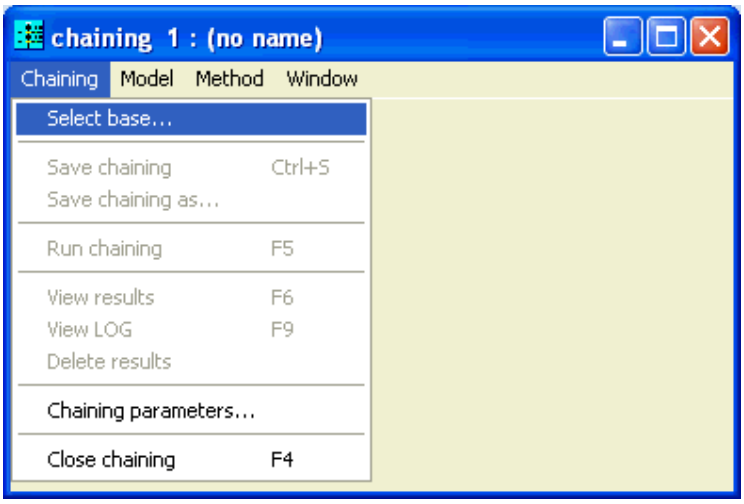


Figura 50 – Selecionando DataBase

Pode-se verificar através da figura 51, que a base foi selecionada com sucesso, sendo exibido ao lado esquerdo do ícone “Base” informações sobre o nome do arquivo “*.sds” selecionado.

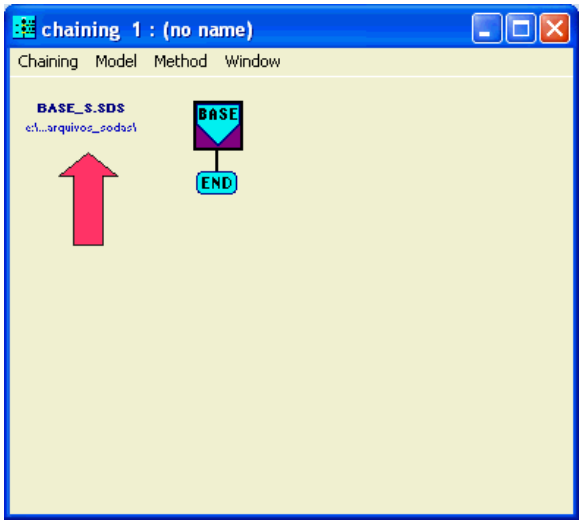


Figura 51 – Base de Dados Selecionada

Podem ser obtidas maiores informações sobre a base utilizada, clicando duas vezes sobre o ícone “Base”, sendo exposta assim informações, como, por exemplo: caminho completo do arquivo, quantidade e tipo de variáveis, objetos simbólicos e título e subtítulo da base. (Fig. 52)

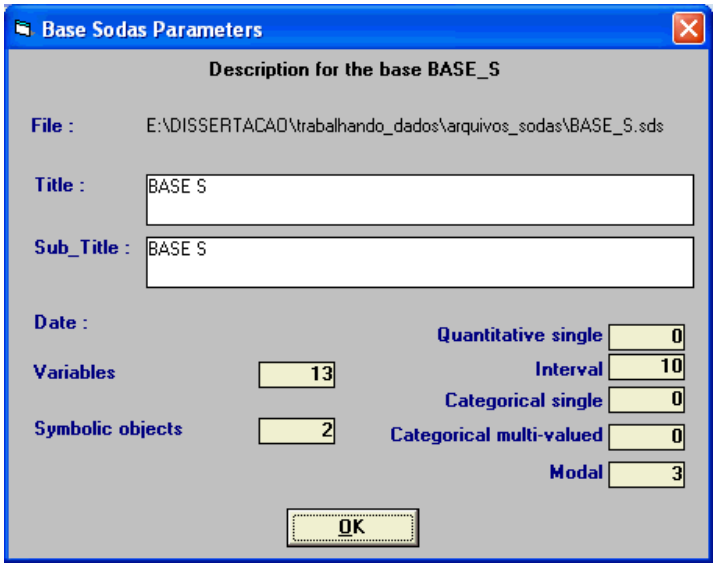


Figura 52 – Informações sobre a Base de Dados

O SODAS utiliza o conceito de "encadeamento" (*chaining*) para os módulos (métodos) utilizados, sendo esta uma outra grande vantagem do *software*, pois facilita a visualização e criação do modelo para o usuário e também possibilita diversas análises em conjunto.

O próximo passo é a inserção de um novo método na cadeia, o que pode ser realizado através da opção "*Method -> Insert Method*" (fig. 53) ou clicando com o botão direito sobre o ícone "Base" e escolhendo a opção "*Insert Method*".

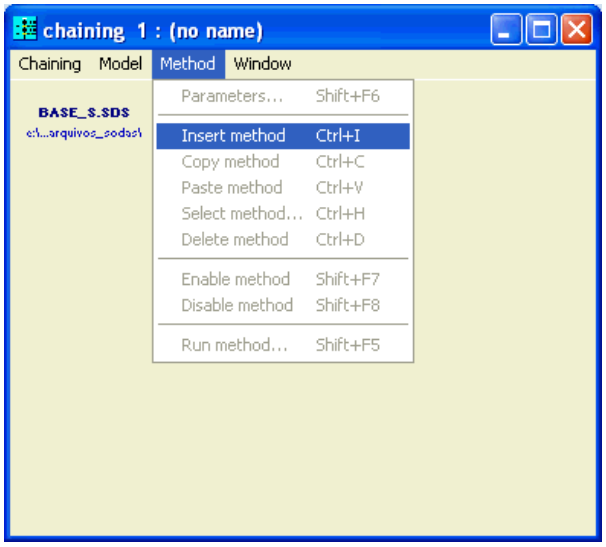


Figura 53 – Inserindo Método

Uma vez inserido o método, deve-se simplesmente clicar sobre o módulo desejado, neste momento o módulo "*View*", e arrastá-lo para o método (Fig. 54) e a caixa do método irá ficar com o nome do módulo escolhido.

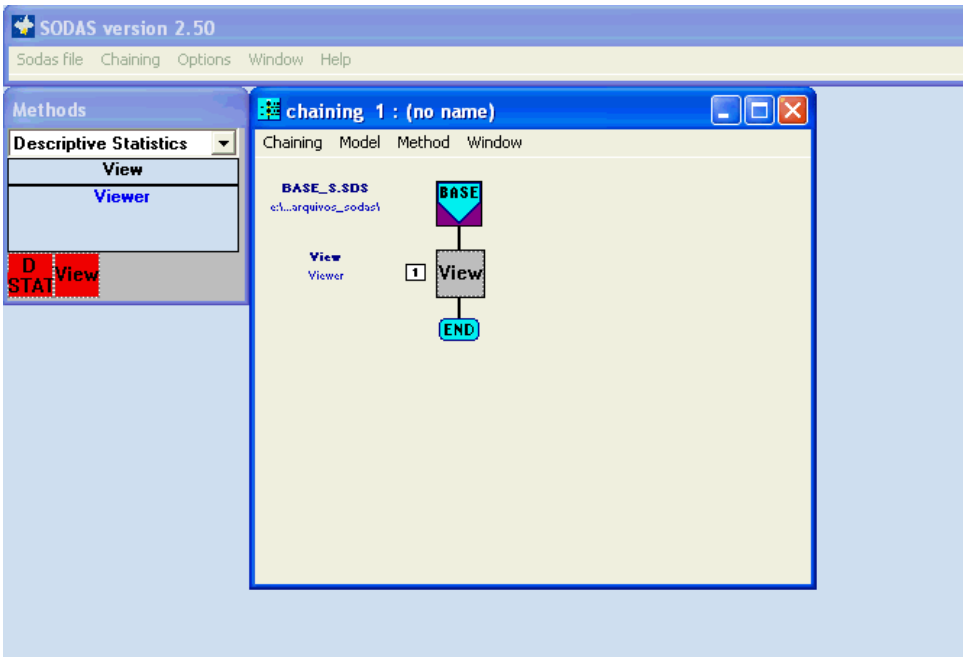


Figura 54 – Método *View* Inserido na cadeia

Para configurar um método, deve-se clicar duas vezes sobre o seu ícone. Sendo assim após efetuar os dois cliques no ícone "*View*" é exibida uma janela com as opções para sua configuração (Fig. 55). Nesta etapa o usuário pode escolher quais variáveis são utilizadas na análise.

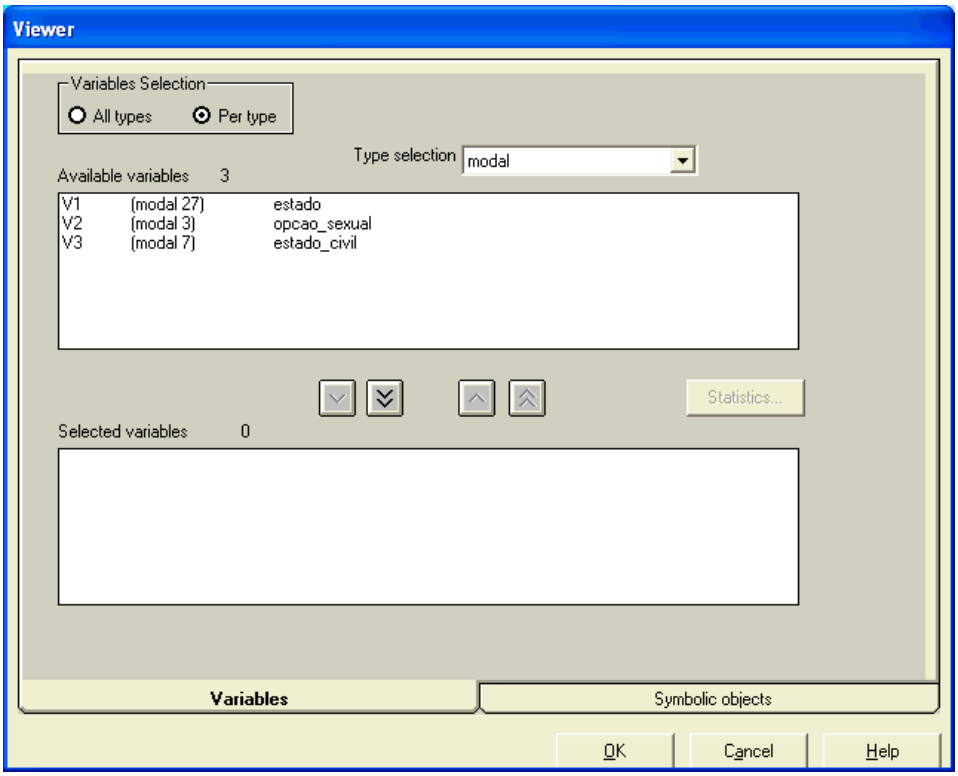


Figura 55 – Configurando o View

Através da aba "*Symbolic Objects*", demonstrada na figura 56, é possível configurar quais Objetos Simbólicos são usados na análise e clicar em "Ok", finalizando assim a configuração do método view.

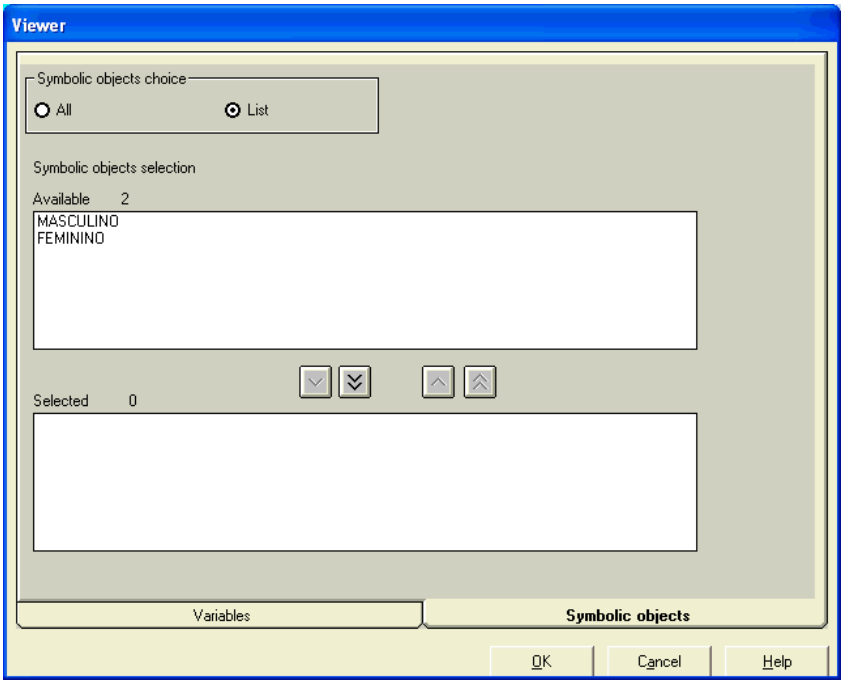


Figura 56 – Seleção de Objetos Simbólicos a serem utilizados pelo *View*

Uma vez configurado o método, deve-se rodá-lo através da opção "*Method -> Run method*" para assim ser gerado o resultado. (Fig. 57)

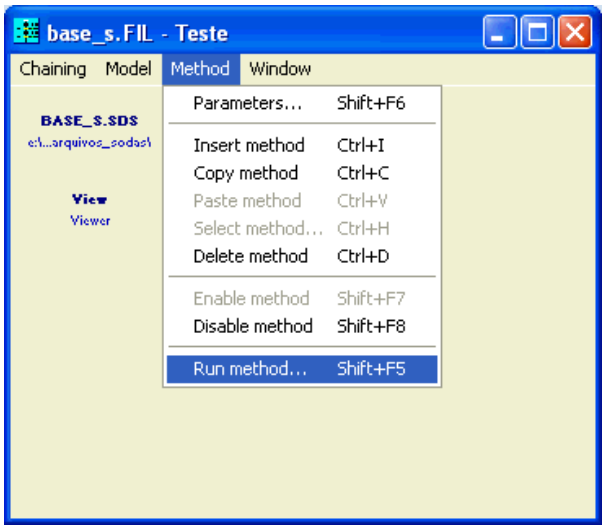


Figura 57 – "Rodando" um método

Uma vez tendo “rodado” o método, o SODAS nos fornece dois *outputs* (Fig. 58) o primeiro sendo somente um *log* (“diário”) da execução do método, o que no caso do *View* somente nos mostra quantos objetos simbólicos foram utilizados. Entretanto para outros métodos que são demonstrados neste trabalho, este *log* contêm informações relevantes. O outro *output* é o resultado em forma gráfica do método View, que pode ser aberto clicando duas vezes sobre o mesmo. (Fig. 59)

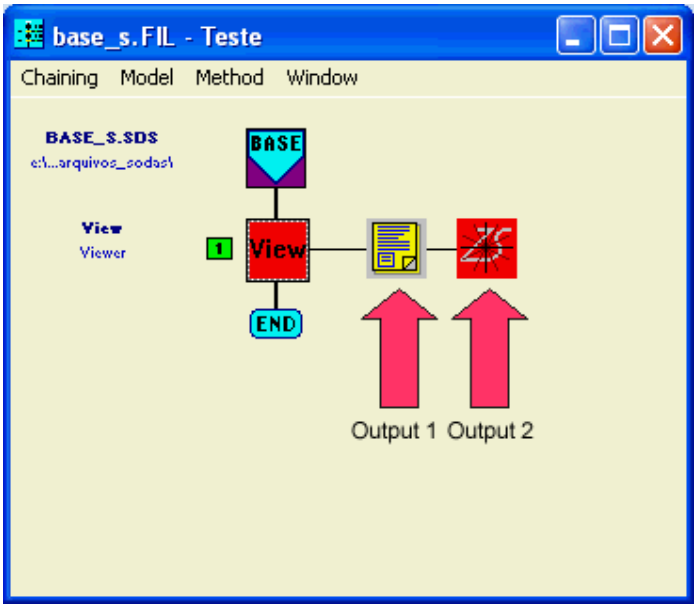


Figura 58 – Método View após ser rodado.

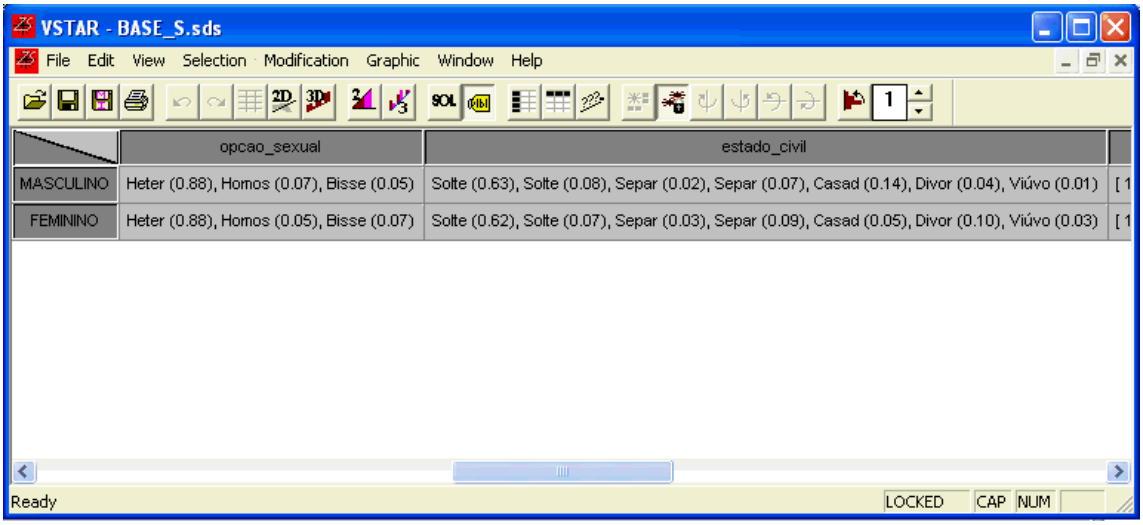


Figura 59 – VSTAR

A equipe do SODAS desenvolveu um aplicativo chamado VSTAR (Fig. 59), que é utilizado pelo método *View* e diversos outros métodos para gerar gráficos e organizar melhor as informações. Este aplicativo, conforme se pode observar já possui em sua abertura algumas informações básicas sobre a base, que são a origem para os gráficos. Por padrão o VSTAR separa os objetos simbólicos em linhas, no exemplo a variável referente ao "sexo" dos usuários, e nas colunas as outras variáveis. Para incluir um objeto simbólico ou uma variável na análise basta clicar sobre seu nome e assim a mesma já constará nos gráficos. Uma boa definição de qual variável será utilizada como objeto simbólico é de grande importância, pois o VSTAR é capaz de gerar gráficos separados de todas variáveis para cada objeto.

O menu do VSTAR é em parte intuitivo, sendo suas funções mais importantes contidas nos ícones:

-  Gráfico em duas dimensões com todas as variáveis selecionadas para cada Objeto simbólico, possibilitando uma análise em 2D para cada objeto. (Fig. 60)
-  Exibe um gráfico em três dimensões com todas as variáveis selecionadas para cada Objeto simbólico, possibilitando uma análise em 3D para cada objeto. (Fig. 61)

-  Gera um gráfico em duas dimensões com todas as variáveis selecionadas, entretanto consolidando as informações de todos Objetos simbólicos. (Fig. 62)
-  Gera um gráfico em três dimensões com todas as variáveis selecionadas, entretanto consolidando as informações de todos Objetos simbólicos. (Fig. 63)
-  Exibe um relatório com todas os dados de cada variável para cada objeto simbólico. (Fig. 64)
-  Quando selecionado exibe os nomes das variáveis propriamente ditos, quando não selecionado, expõem o nome que o SODAS utiliza para as variáveis. Pode ser percebida a diferença na figura 65, onde se utiliza a opção do link SOL, entretanto com o ícone “label” não selecionado.

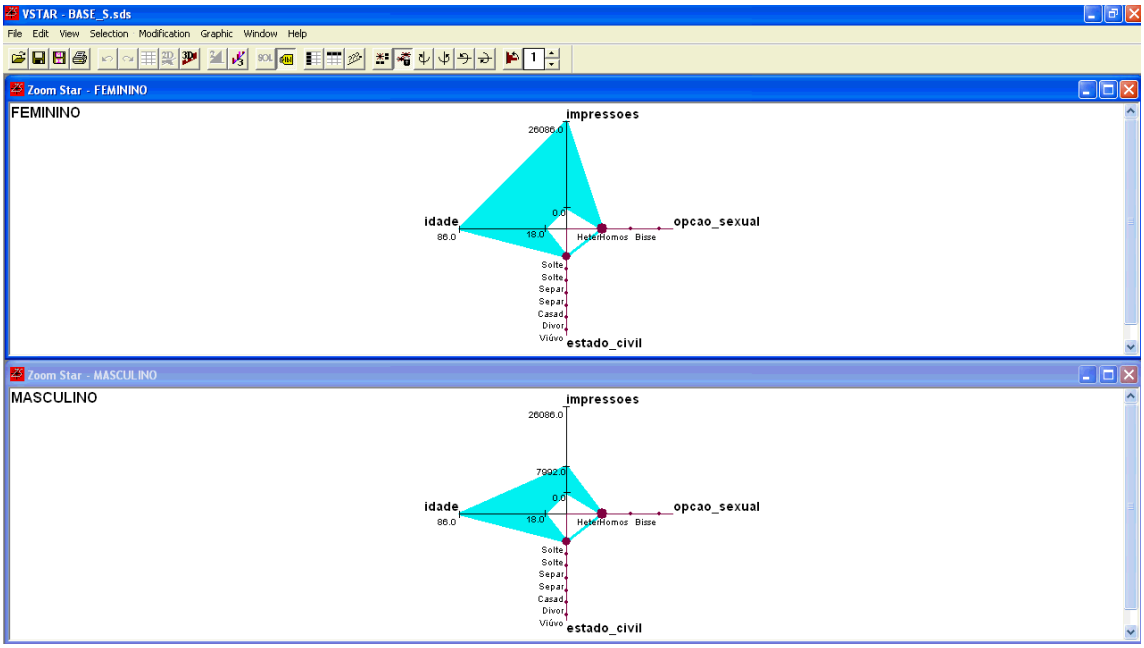


Figura 60 – 2D Gráficos Separados

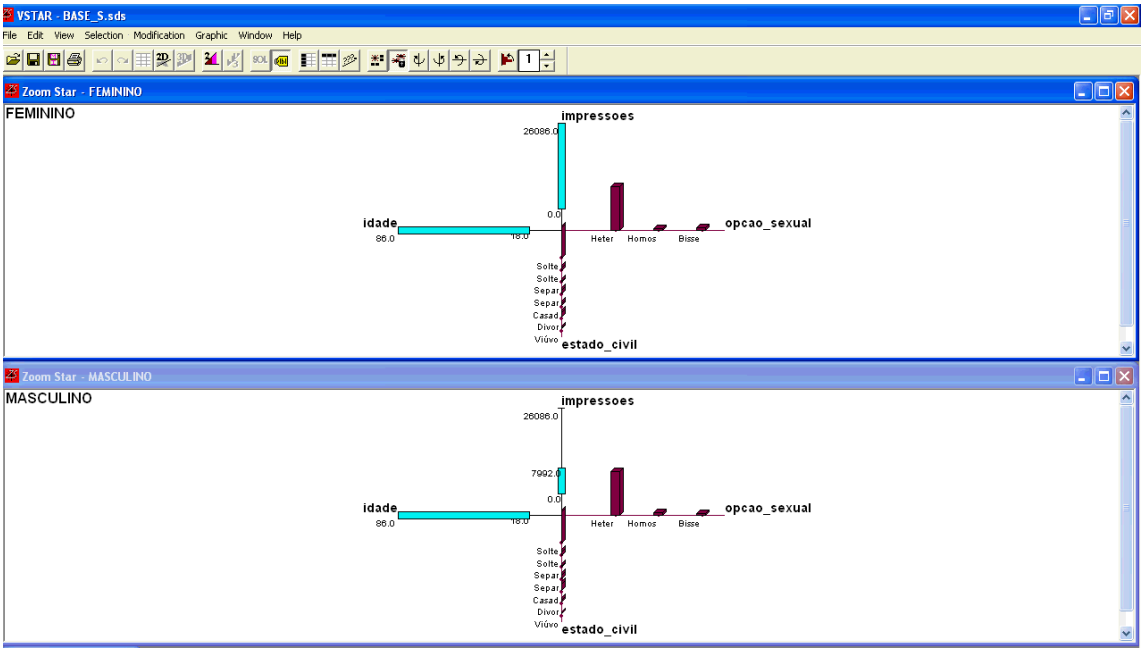


Figura 61 – 3d gráficos separados

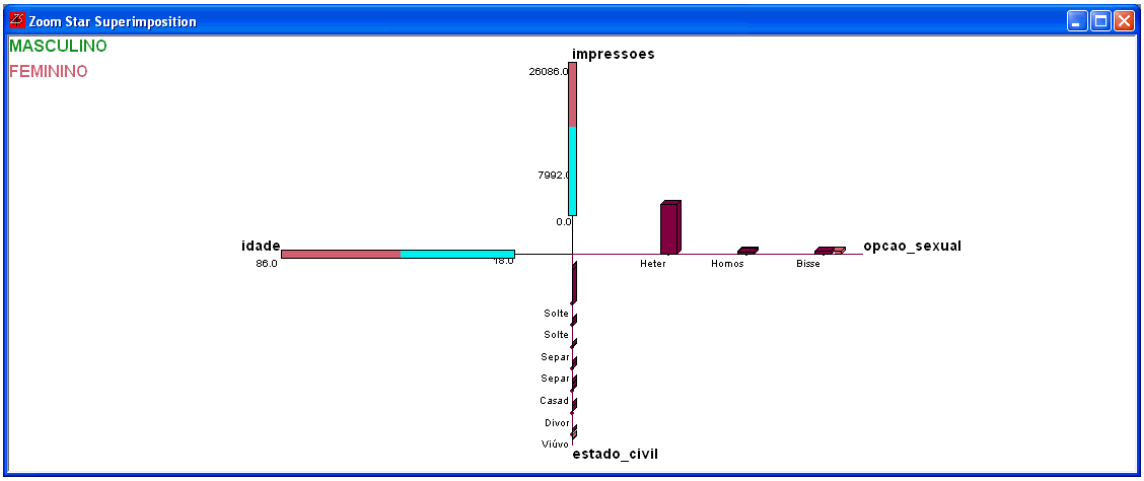


Figura 62 – Gráfico 2d consolidado

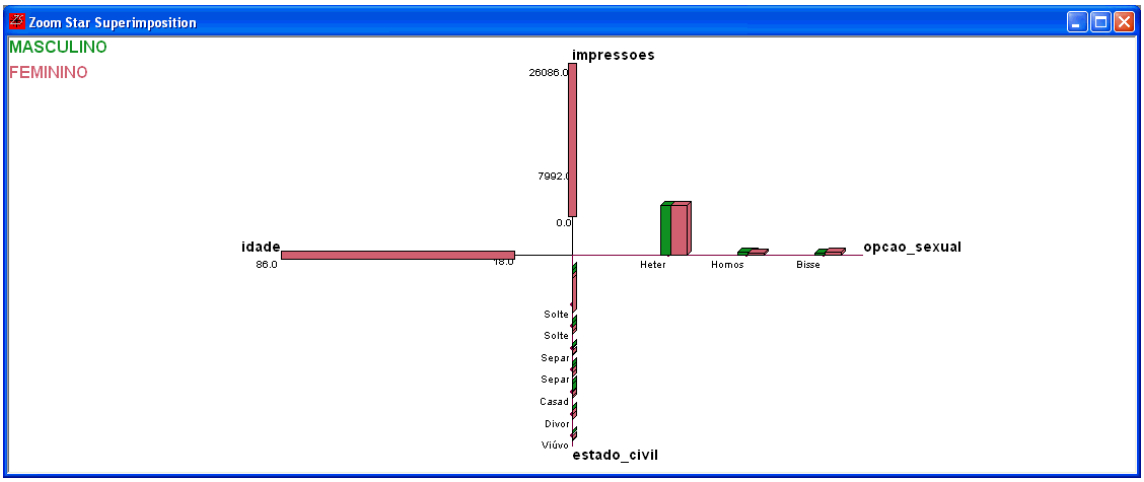


Figura 63 – 3d consolidado

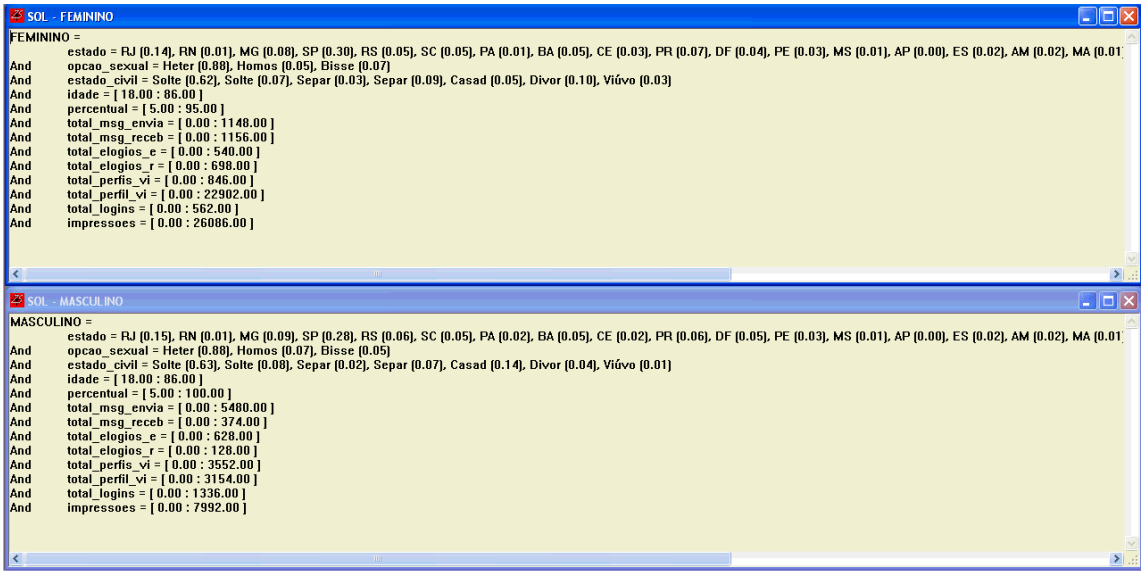


Figura 64 – SOL label selecionado

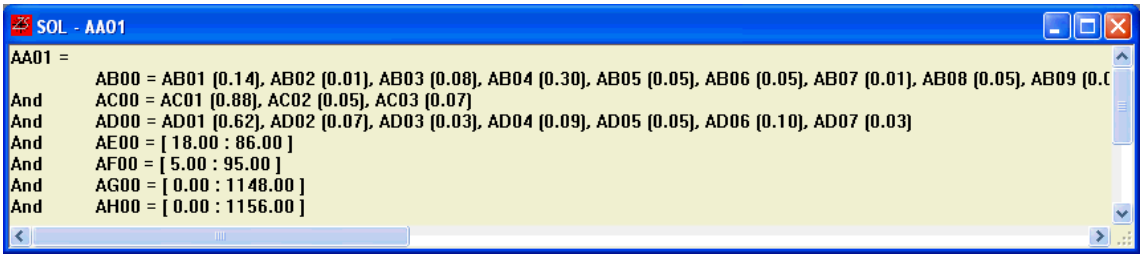


Figura 65 – SOL label não selecionado (sexo Feminino)

DSTAT

Nesta etapa é descrito o método DSTAT, que juntamente com o método View gera a estatística descritiva básica da base de dados, iniciando assim todo processo de *Data Mining*.

O método DSTAT é mais completo que o View, oferecendo mais informações, e também gráficos mais específicos.

Inicia-se este passo a passo a partir de uma base de dados já selecionada, tendo em vista que o processo para selecionar uma base é o mesmo já explicitado no exemplo do método View, sendo necessário no primeiro momento arrastar o método DSTAT para a cadeia. (Fig. 66)

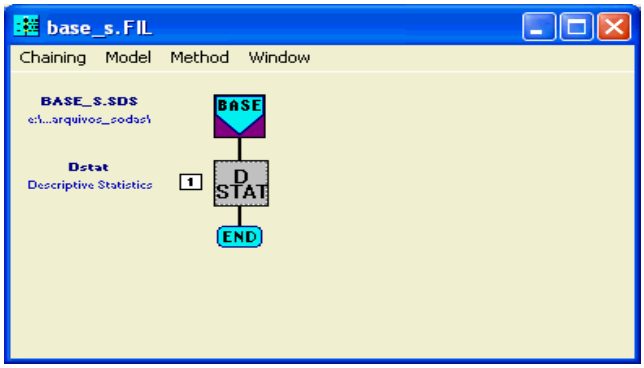


Figura 66 – Método DSTAT incluso na cadeia

Após incluir o DSTAT, deve-se clicar duas vezes sobre o seu ícone, e assim é exposta a sua janela de configuração, conforme figura 67.

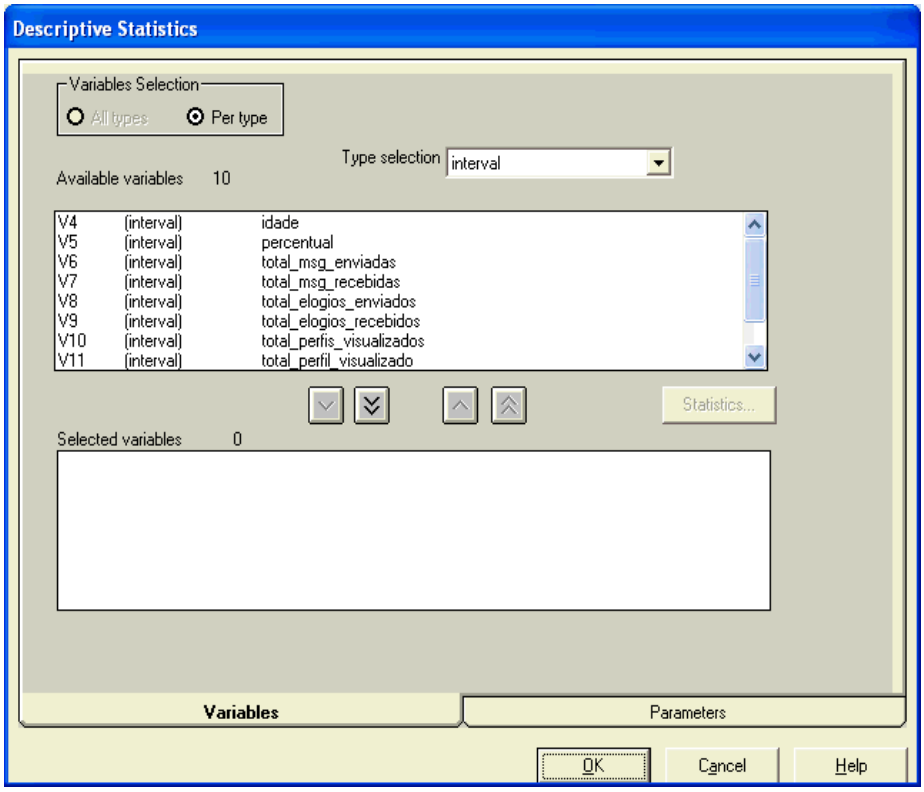


Figura 67 – Configuração de Variáveis no módulo DSTAT

Nesta etapa deve-se escolher as variáveis que se deseja analisar. Entretanto, no caso do método DSTAT, não se pode efetuar uma mistura de variáveis, por exemplo, não é possível escolher variáveis "*Modal*" e ao mesmo tempo do tipo "*Interval*". Isto ocorre, tendo em vista que este método trata de forma diferente cada tipo de variável. Depois de selecionadas as variáveis desejadas, deve-se acessar a aba "*Parameters*", onde é efetuada a escolha do método de análise utilizado de acordo com o tipo de variável escolhida na etapa anterior.

Conforme abordado anteriormente, no estudo de caso, são utilizados variáveis do tipo *Modal* e *Interval*. Sendo assim na configuração dos "*Parameters*", deve-se selecionar o método de análise de acordo com o tipo de variável escolhido. Pode-se verificar na tabela 6, quais opções estão disponíveis para cada tipo de variável.

	Interval	***Modal***
Frequencies for categorical multi-values		
Frequencies for *Interval*	X	
Biplot	X	
Capacities		X
Numeric and symbolic characteristics	X	

Tabela 6 – Métodos disponíveis de análise do DSTAT por tipo de variável.

São utilizadas neste primeiro exemplo variáveis do tipo *Modal* e método "*Capacities for Modal variables*" (Fig. 68). Após configurar as variáveis, deve-se rodar o método para obter o seu output (Fig. 69).

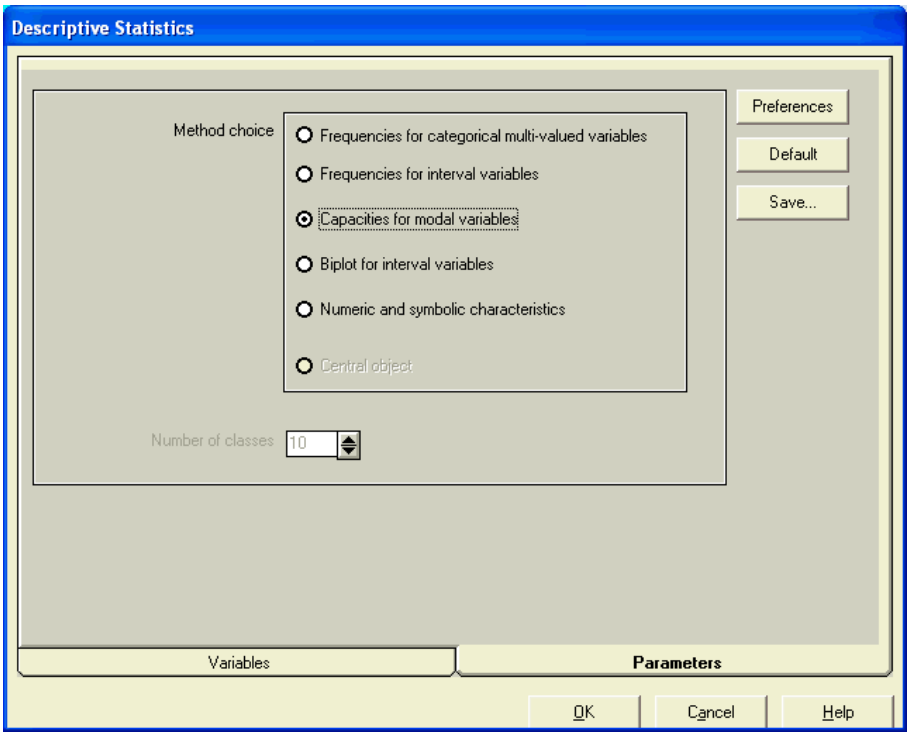


Figura 68 – Configuração parâmetros do DSTAT para variável *Modal*

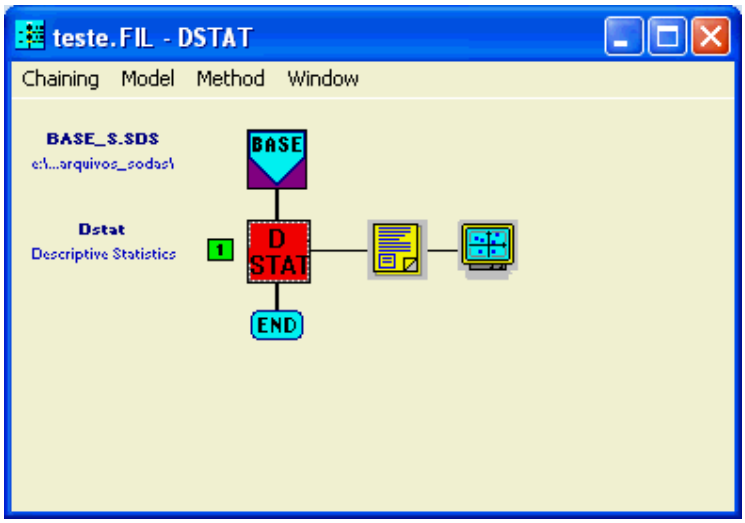


Figura 69 – *Output* DSTAT após rodar o Método

O formato do *output* é semelhante ao utilizado no método *View*, gerando um resultado em texto e um resultado gráfico. Clicando duas vezes sobre o primeiro *output*, são verificadas as informações em formato de texto relevantes, sendo os valores "mínimo, máximo e média" isto quando se esta analisando variáveis "*Modal*" (Fig. 70). Os valores gerados são baseados na análise por Objeto Simbólico, neste caso a variável "sexo" (Masculino ou Feminino). Sendo assim pode-se verificar que o estado do São Paulo (SP), por exemplo, possui uma participação máxima no valor 0,2969 (29,69%) em um determinado sexo e uma participação mínima de 0,2777 (27,77%) em outro sexo e uma média de 0,2873 (28,73%).

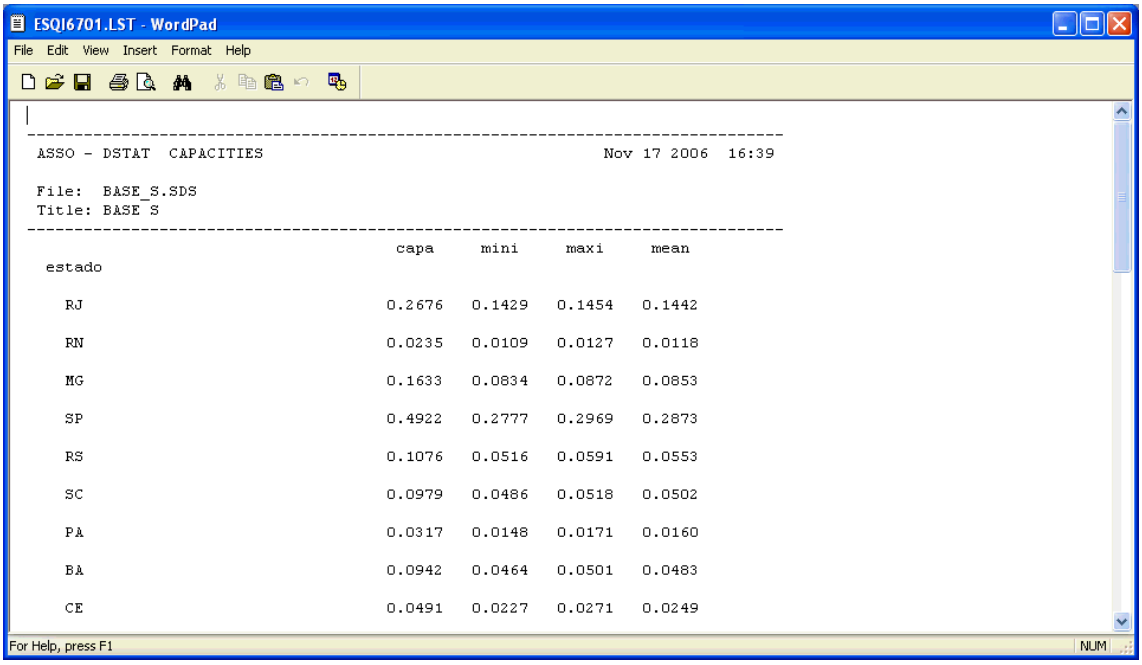


Figura 70 – *Output* 1 DSTAT (*Capacities for Modal*)

Acessando os resultados do segundo *output*, são obtidos graficamente os resultados apresentados no formato de texto, facilitando assim a interpretação e comparação das informações. Inicialmente é exposta uma janela para seleção de Variáveis as quais deseja-se obter gráficos, podendo ser escolhida mais de uma variável ao mesmo tempo. (Fig. 71)

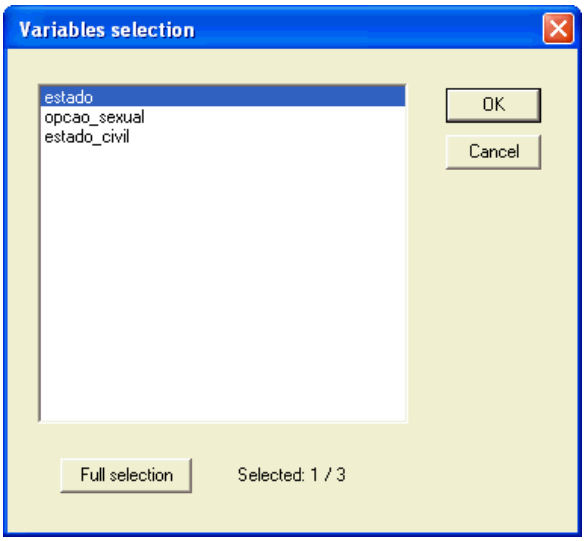


Figura 71 – *Output 2* DSTAT (*Capacities for Modal*) – Selação de Variáveis

Após efetuar a seleção das variáveis e clicar no botão de “Ok”, é exposto o gráfico desejado, figura 72.

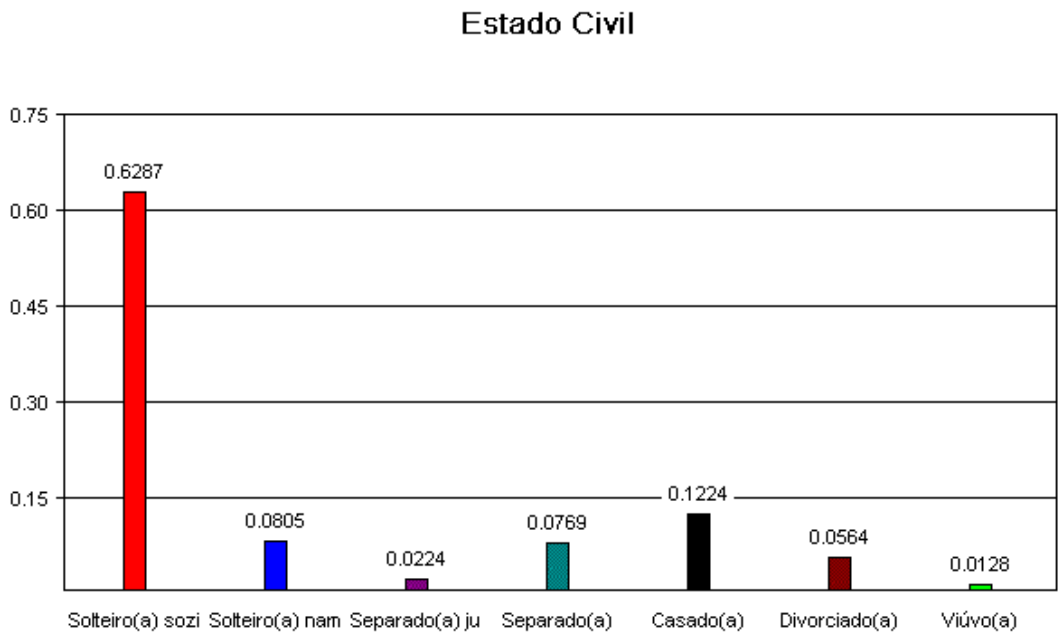


Figura 72 – *Output 2* DSTAT (*Capacities for Modal*) – Gráficos

É possível também gerar um gráfico com a Média, Mínimo e Máximo separados, para obter esta informação, deve-se clicar no ícone  disponível no menu. (Fig. 73)

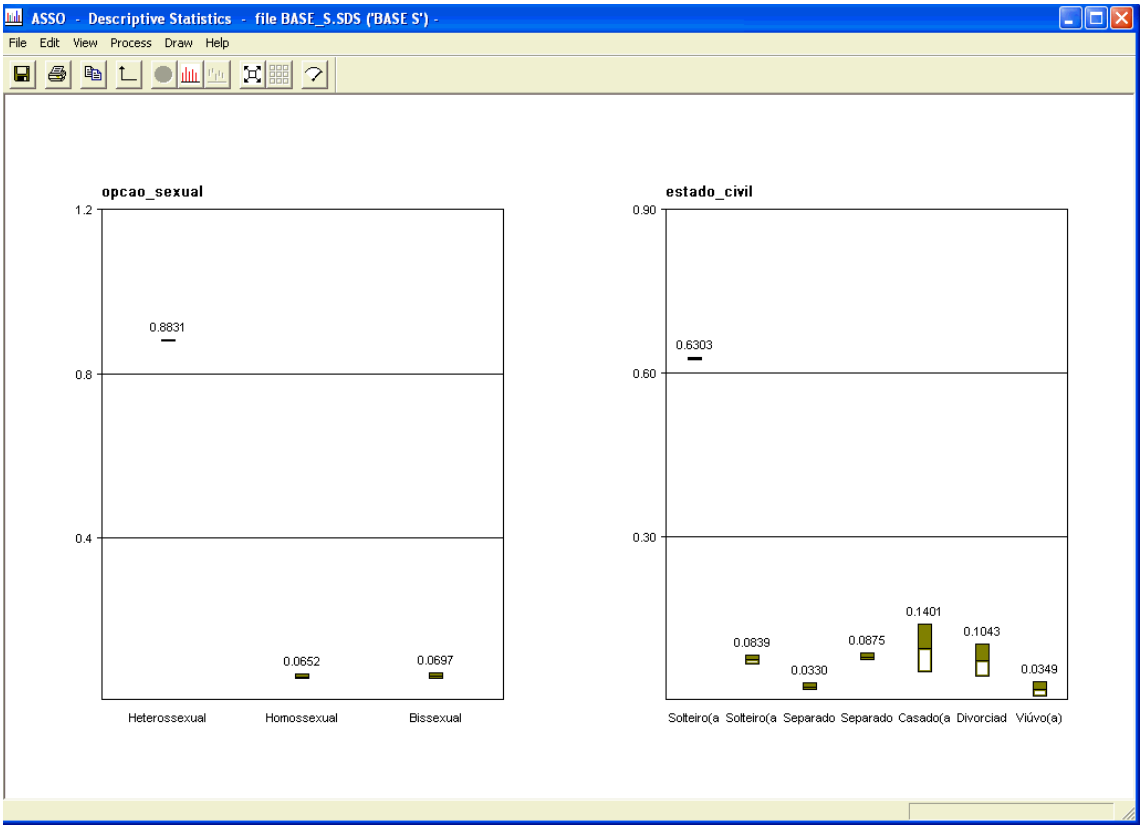


Figura 73 – *Output 2* DSTAT (*Capacities for Modal*) – Gráficos (min., Max. e média)

O módulo DSTAT permite somente este teste para variáveis "*Modal*". Sendo assim, agora são utilizadas variáveis do tipo "*Interval*", onde se têm maiores opções. Deve-se então, novamente, clicar duas vezes sobre o ícone do módulo DSTAT e abrindo assim a janela para seleção de variáveis e configuração dos métodos. Pode ser visualizado na figura 74 a seleção das variáveis do tipo "Interval" utilizadas neste exemplo.

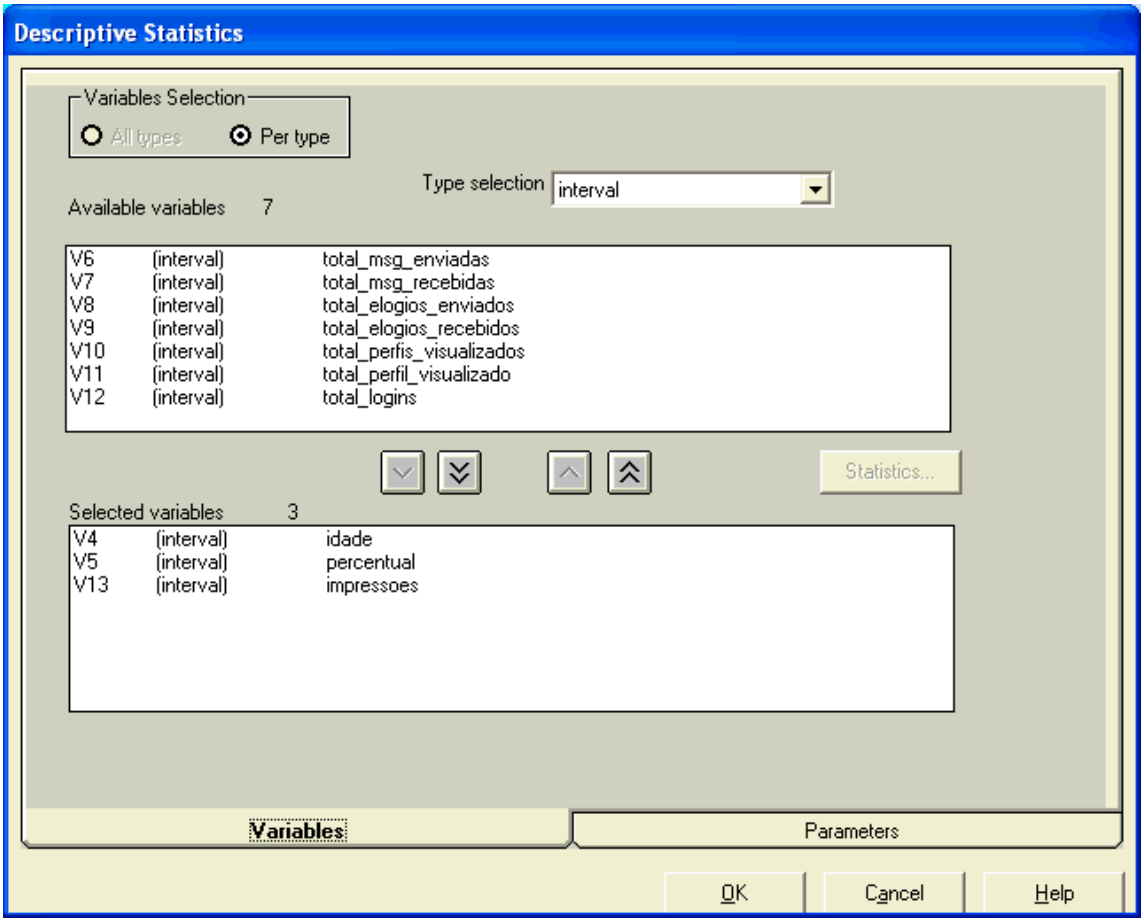


Figura 74 – DSTAT –Selecionando Variáveis *Interval*

Após selecionar as variáveis, deve-se selecionar o método utilizado, onde neste momento se utiliza o método "*Frequencies for Interval variables*". (Fig. 75)

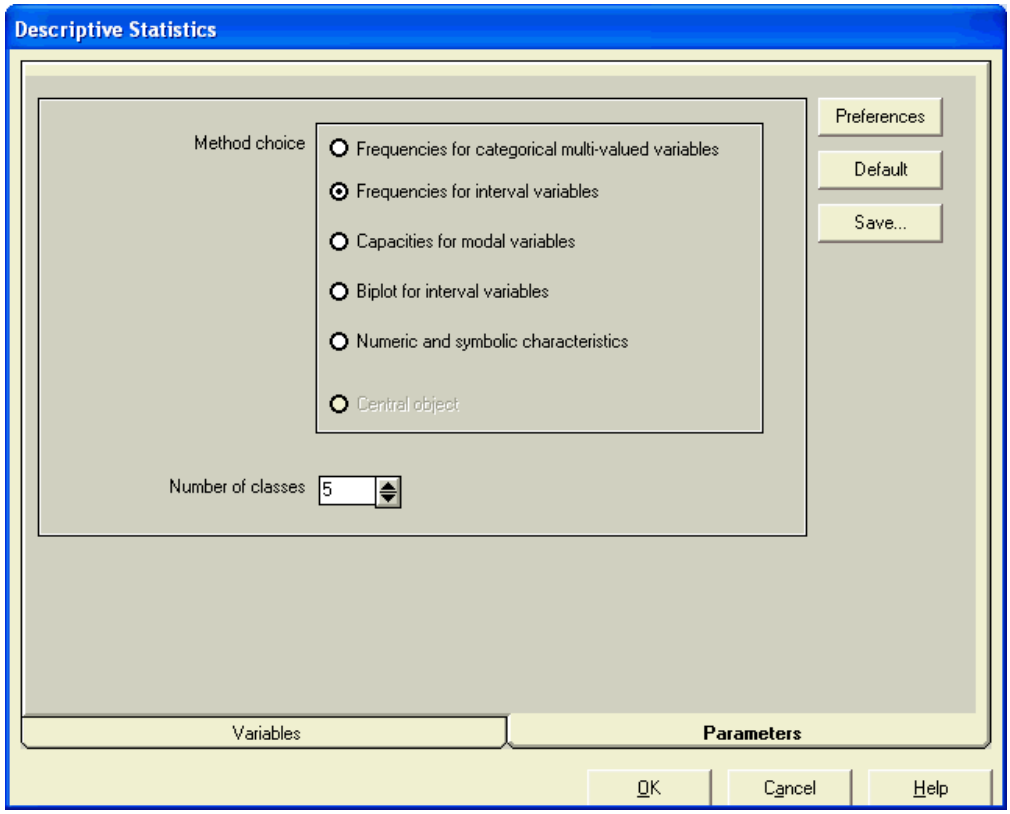


Figura 75 – DSTAT selecionando Parâmetro *"Frequencies for Interval variables"*

Novamente após executar o método, têm-se dois *outputs*, sendo o primeiro no formato texto representado pela figura 76, onde podem ser verificadas informações sobre os limites das variáveis, o tamanho das classes em que foi dividida para análise, a tendência de centro e a dispersão.

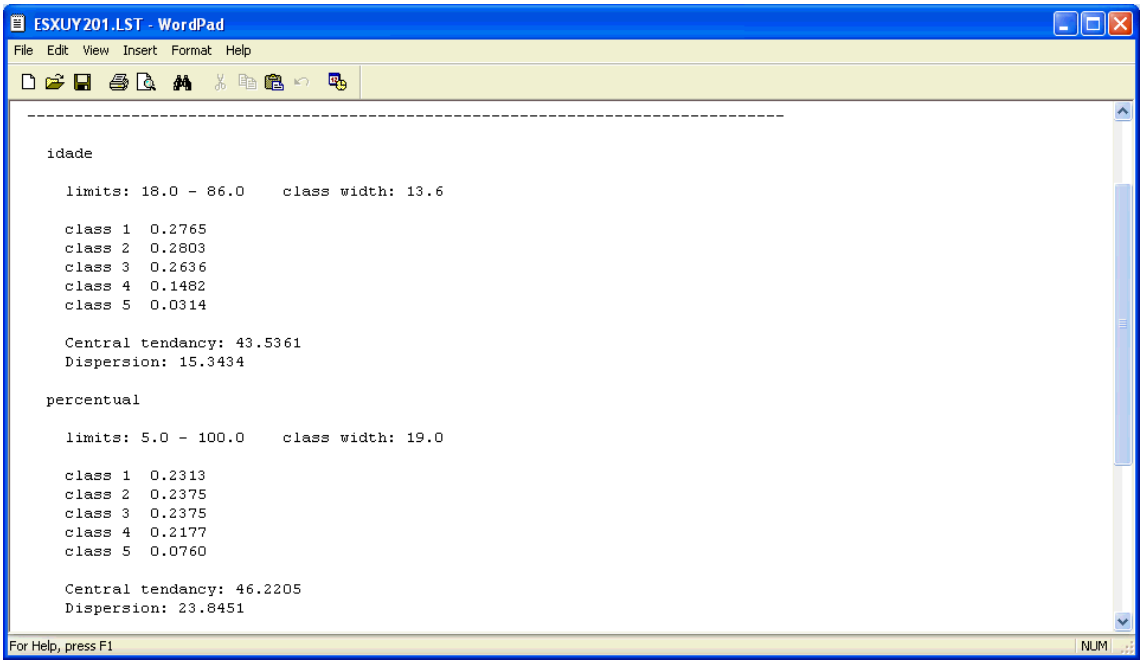


Figura 76 – DSTAT *"Frequencies for Interval variables" output 1*

O Segundo *output* fornece uma representação gráfica da distribuição em percentual para cada tipo de classe, conforme mostrado na figura 77.

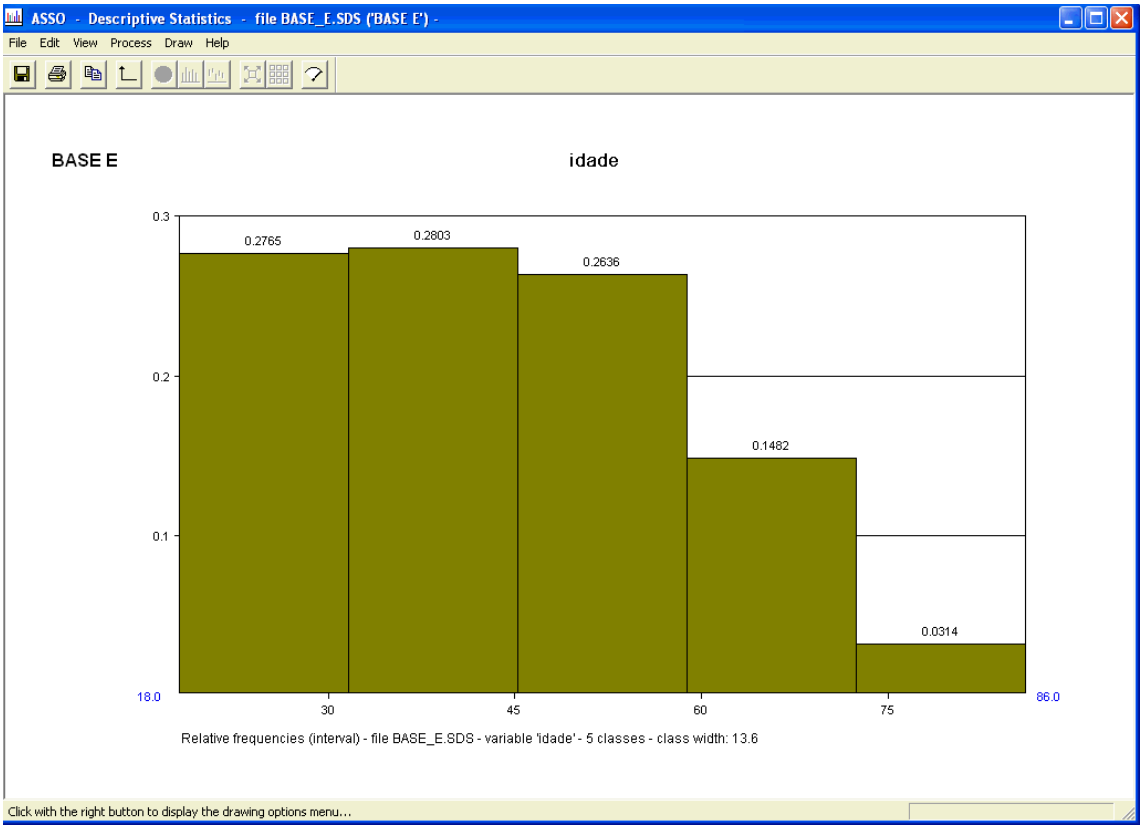


Figura 77 – DSTAT "*Frequencies for Interval variables" output 2*

Outro método muito relevante do módulo DSTAT com variáveis "*Interval*" é o chamado "*Biplot*". Para configurá-lo, deve-se clicar duas vezes sobre o ícone DSTAT e na aba "*Parameters*" selecionar a opção "*Biplot for Interval variables*", conforme demonstrado na figura 78.

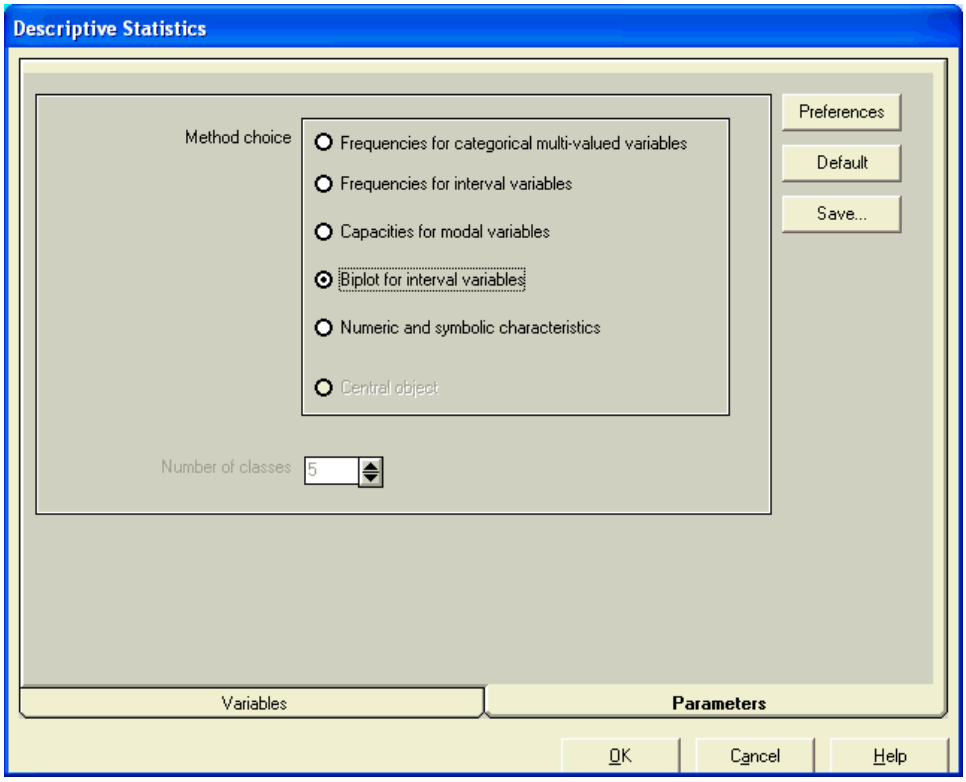


Figura 78 – DSTAT selecionando Parâmetro "*Biplot for Interval*"

O método *Biplot* também oferece dois *outputs*. O primeiro torna-se irrelevante, pois é somente um *log* simples, com data e nome do arquivo executado. Entretanto, quando ocorre algum erro, este *log* pode ser interessante para descobrir falhas. Sendo assim o segundo *output* demonstra graficamente o resultado. É preciso primeiramente selecionar as duas variáveis necessárias para gerar o gráfico *Biplot*, figura 79.

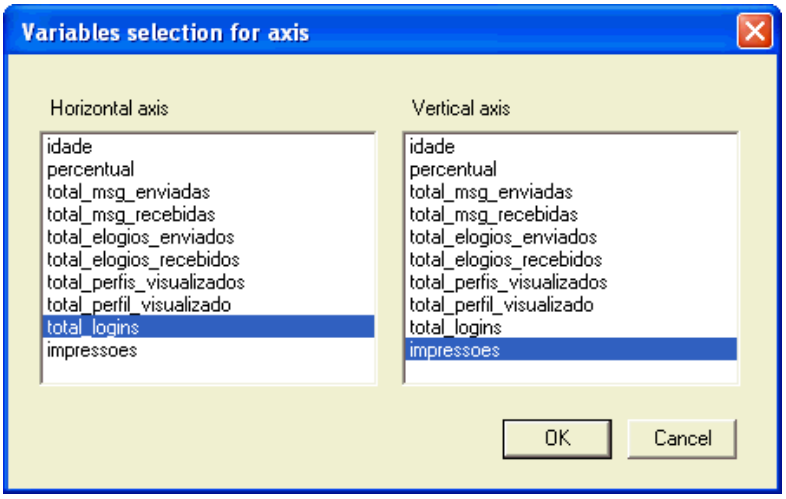


Figura 79 – DSTAT Parâmetro *"Biplot for Interval" output 2*

Após selecionar as duas variáveis e clicar no botão de "Ok", é mostrado o gráfico desejado com as duas variáveis escolhidas, neste caso "impressões" e "total_logins", e também divisões dentro do gráfico de acordo com os Objetos Simbólicos, onde neste exemplo é utilizada a variável "Estado". (Fig. 80)

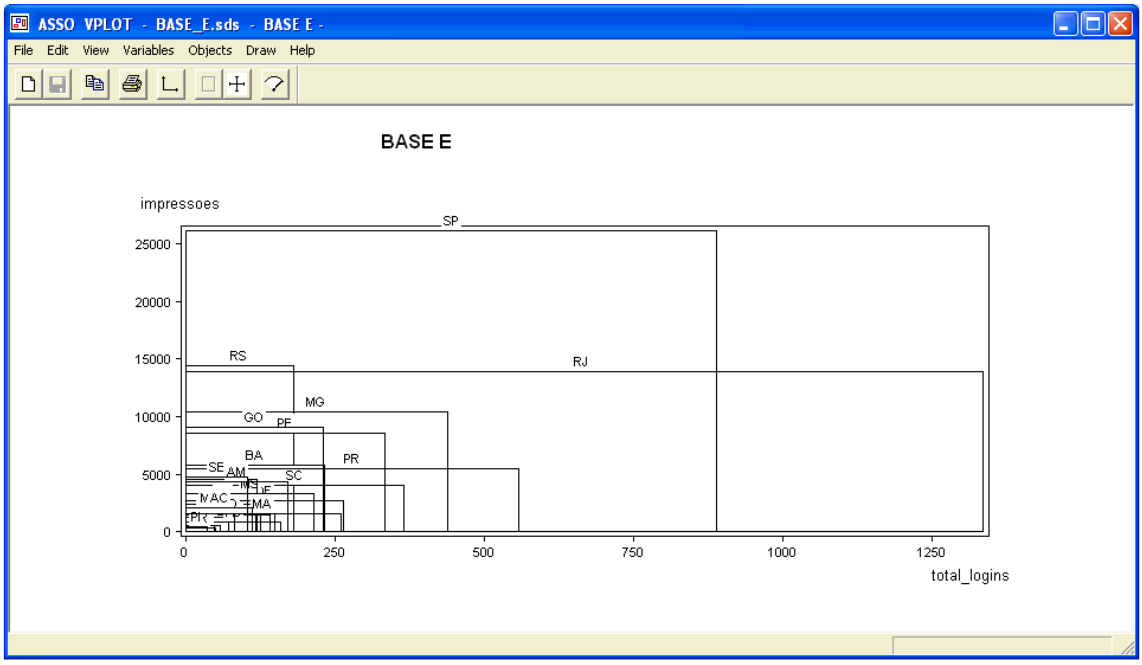


Figura 80 – DSTAT Parâmetro *"Biplot for Interval" output gráfico*

O último exemplo de utilização do módulo DSTAT usando variáveis *Interval* é o método "*Numeric and symbolic characteristics*". Para utilizar este método, deve-se clicar novamente duas vezes no ícone DSTAT e selecionar a referida opção, conforme figura 81.

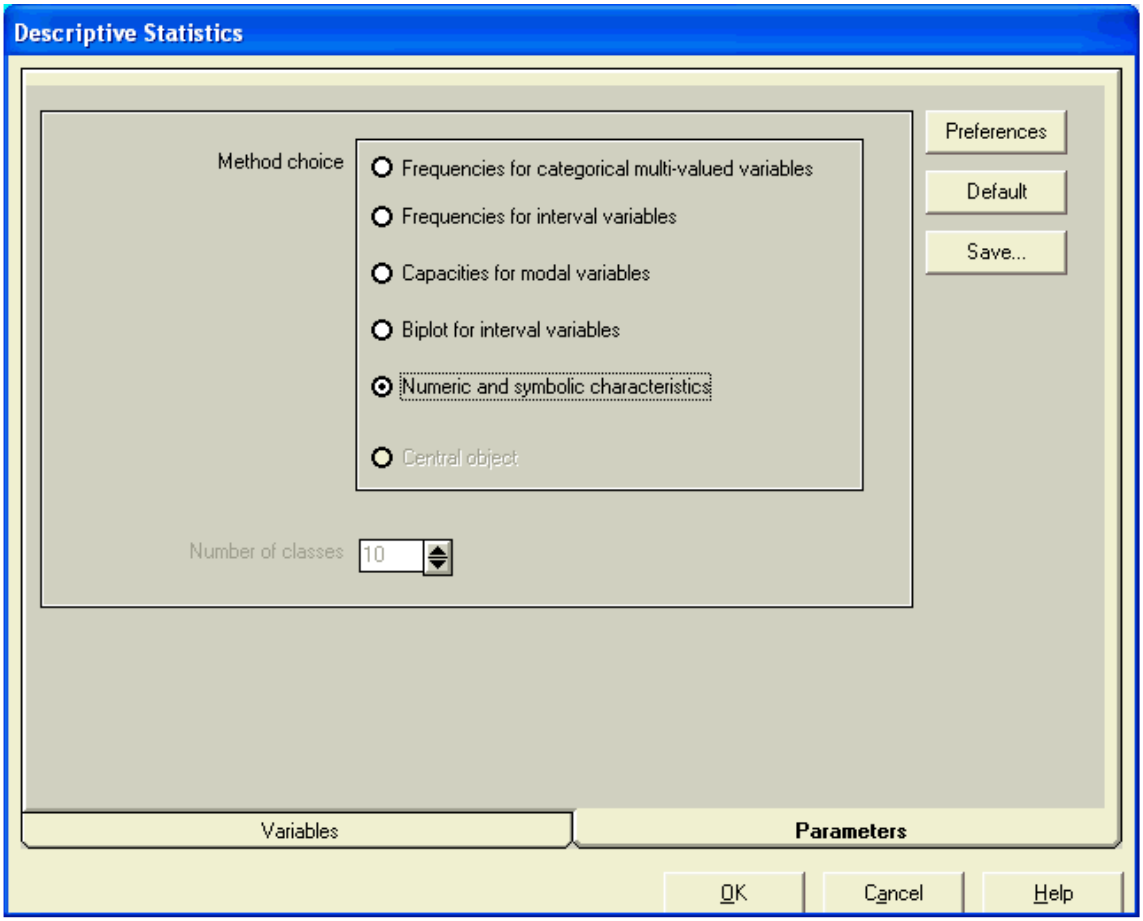


Figura 81 – DSTAT selecionando Parâmetro "*Numeric and symbolic characteristics*"

Este método possui somente uma forma de *output*, onde após rodá-lo, é aberta uma janela (Fig. 82), onde é possível cruzar informações de duas variáveis, sendo exposta à média e desvio padrão de cada uma e a correlação entre as mesmas. Por padrão a janela é aberta com a opção "*Numeric*" selecionada. Sendo assim, os valores mostrados são consolidados, não retratando diferenças de valores por Objetos Simbólicos. Caso se opte pelo método "*Symbolic*", têm-se os limites das médias e do desvio padrão das variáveis escolhidas (Fig. 83).

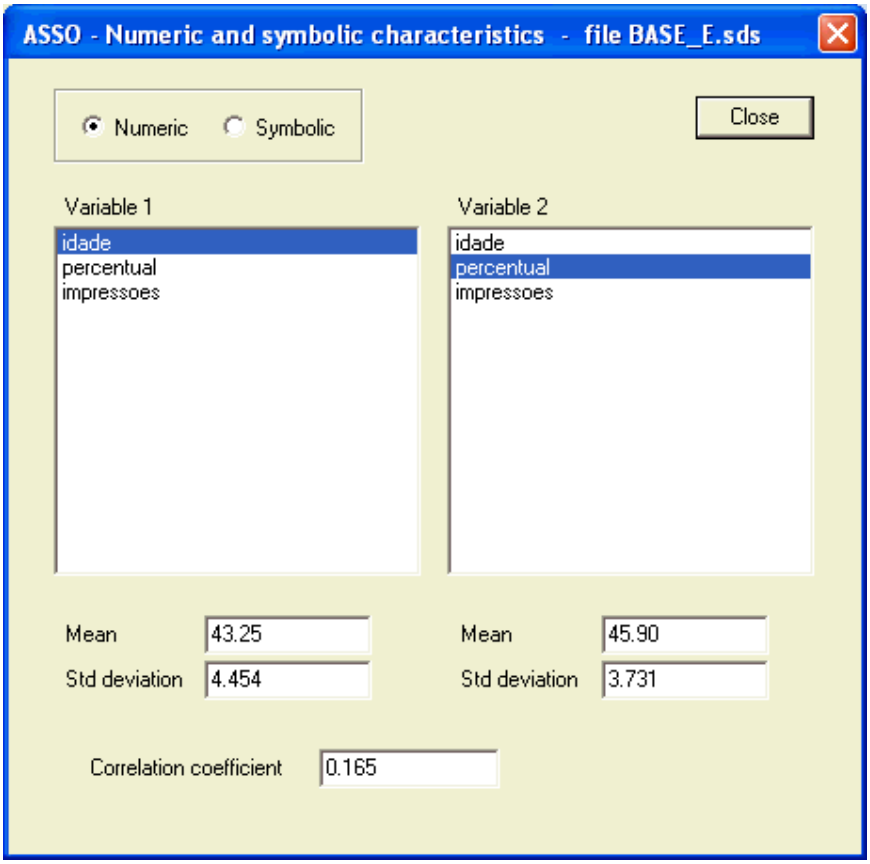


Figura 82 – DSTAT Parâmetro "*Numeric and symbolic characteristics*" *output Numeric*

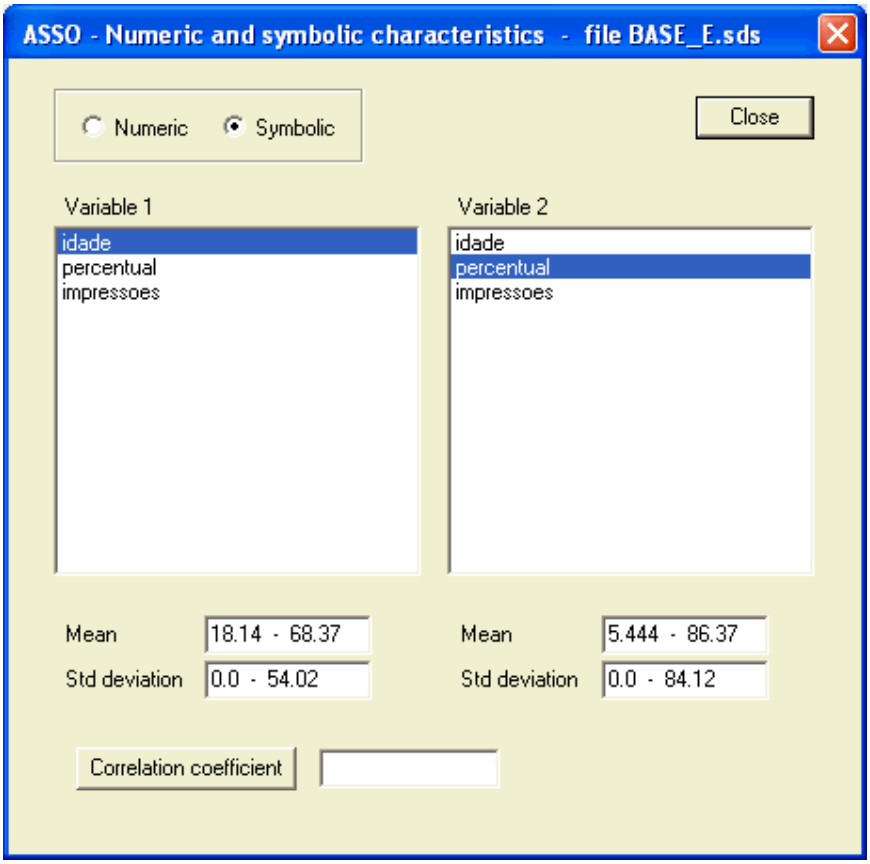


Figura 83 – DSTAT Parâmetro "*Numeric and symbolic characteristics" output Symbolic*

MÓDULO DE CLUSTERING

O SODAS por ser um *software* de *Data Mining*, possui diversos métodos relacionados a *Clustering*, os quais nesta etapa é descrito passo a passo como utiliza-los.

DIV

É mostrado a seguir um exemplo passo a passo da utilização do DIV com variáveis do tipo *Interval*. Deve-se inicialmente arrastar o método DIV para dentro da cadeia. (Fig. 84).

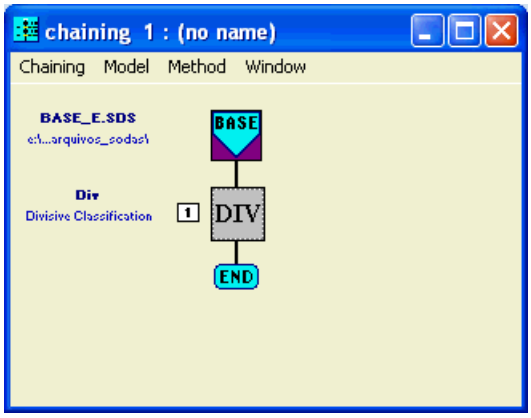


Figura 84 – Método DIV incluso na cadeia

Após inserir o método, deve-se dar um duplo clique sobre o ícone DIV, onde é aberta a janela para seleção das variáveis que serão utilizadas no modelo. O usuário poderá selecionar tanto variáveis qualitativas quanto quantitativas, entretanto não poderá misturá-las no modelo. (Fig. 85).

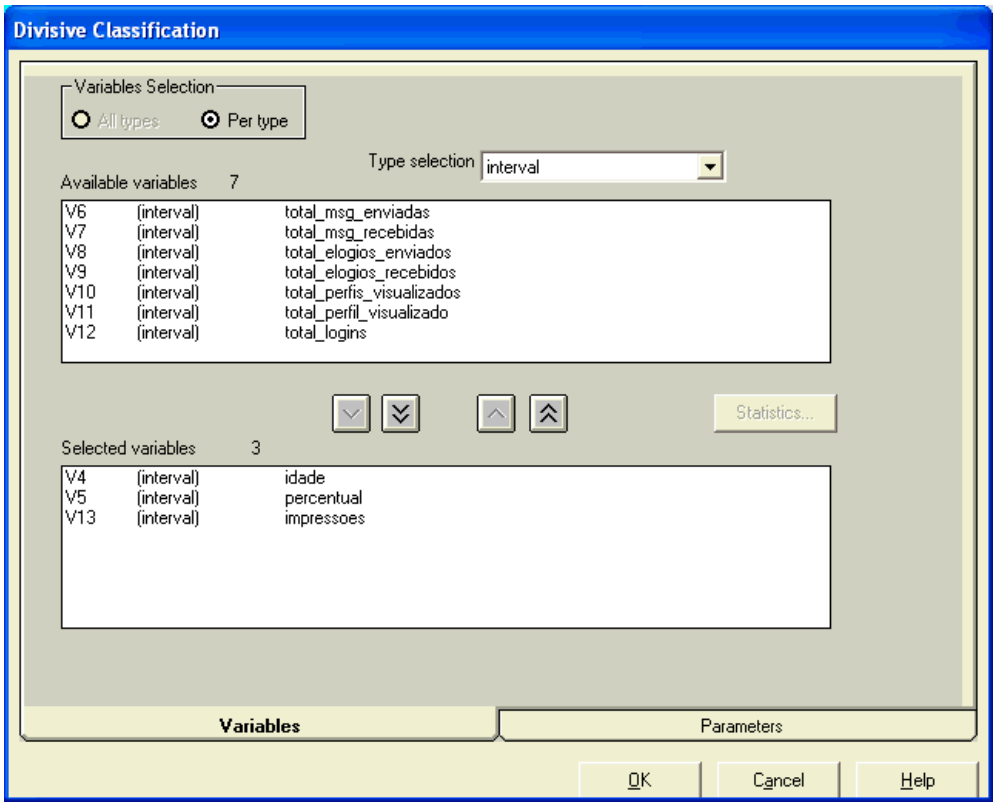


Figura 85 – Método DIV - Selecionando as variáveis

Após realizada a seleção de variáveis, deve-se configurar os parâmetros na aba "*Parameters*", onde é preciso escolher o tipo de normalização e o número de classes para o *cluster*. Caso o usuário queira um *output* extra, poderá também selecionar as opções "*Save partition in Sodas file" e/ou "Save Node Base"*. (Fig. 86)

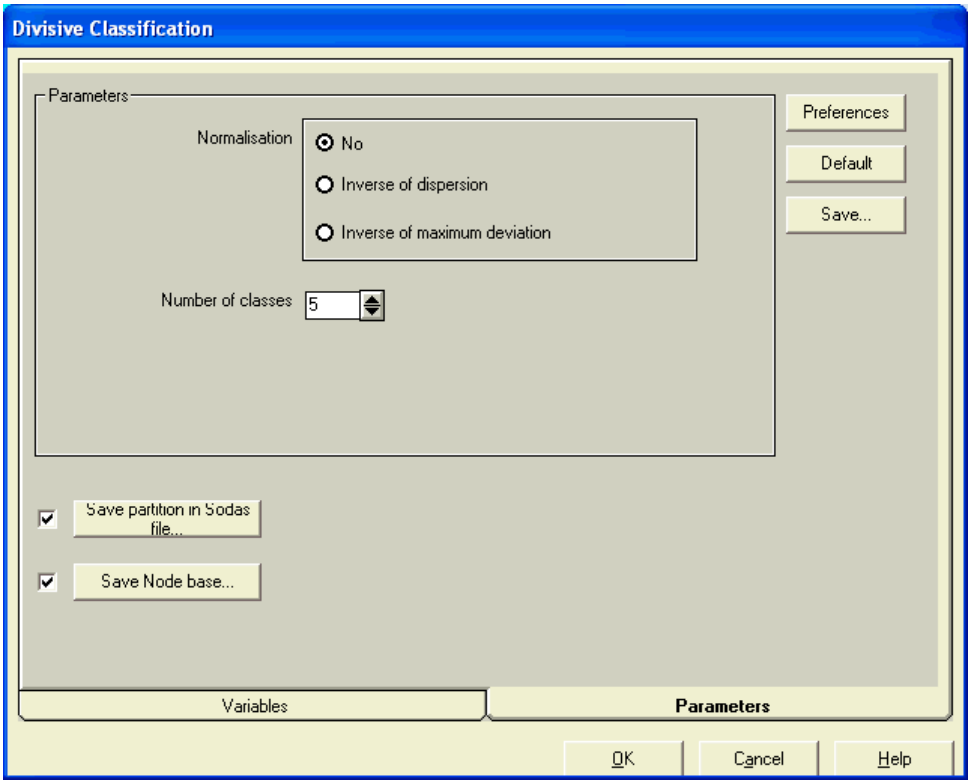


Figura 86 – Método DIV ajustando parâmetros

Uma vez efetuada a configuração desejada do DIV, deve-se rodar o método, gerando assim os *outputs* demonstrados na figura 87.

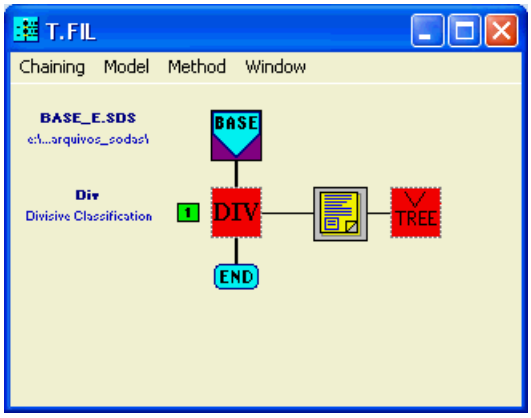


Figura 87 – Método DIV – *outputs gerados*

O método DIV possui por padrão dois *outputs*, sendo o primeiro um arquivo texto e o segundo as informações consolidadas para serem visualizadas no formato de "árvore". Caso se tenha selecionado as opções de *outputs* extras, pode-se também visualiza-los através do método "*View*".

Através do primeiro *output* (Fig. 88), pode-se verificar passo a passo as divisões efetuadas até que seja atingido o número de *clusters* configurado, expondo no mesmo arquivo informações como, por exemplo, o valor de corte e os objetos simbólicos de cada *cluster.*

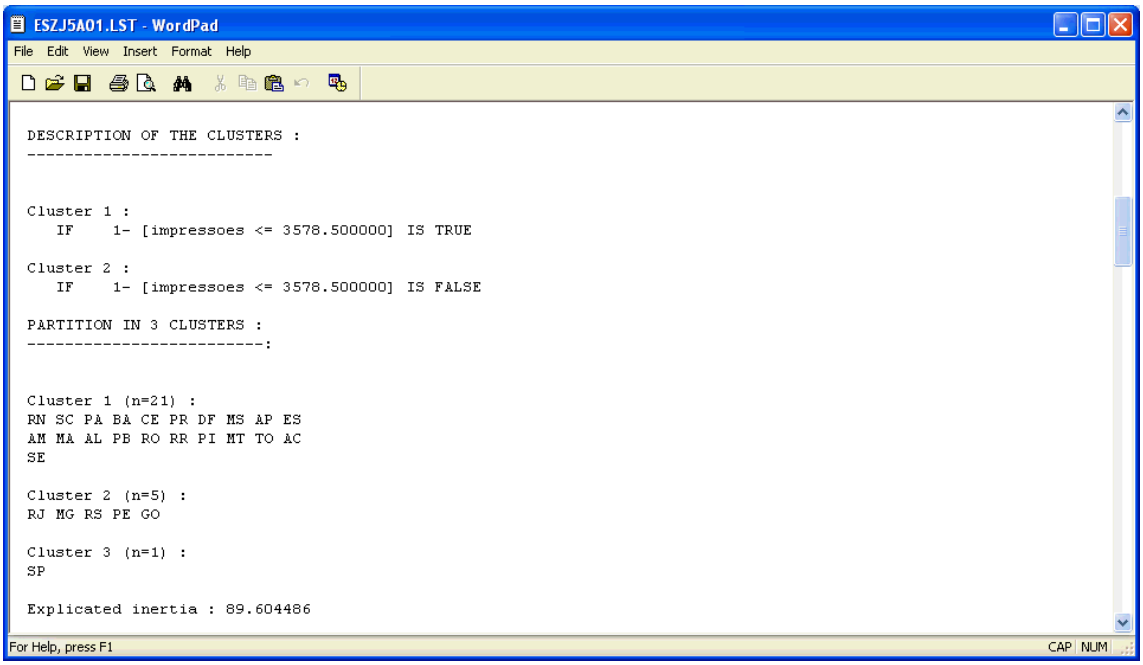


Figura 88 – Método DIV – *output 1* em arquivo texto

O segundo *output* (Fig. 89) demonstra para o usuário o cluster dividido no formato de uma "árvore", onde pode ser verificado claramente como foi a evolução da divisão do *cluster*, tal como a quantidade de objetos simbólicos, a variável de corte e o valor de corte do modelo, melhorando assim drasticamente a facilidade de análise do modelo, uma vez que fica bem claro para o usuário se ele deve acrescentar ou diminuir a quantidade de *clusters.*

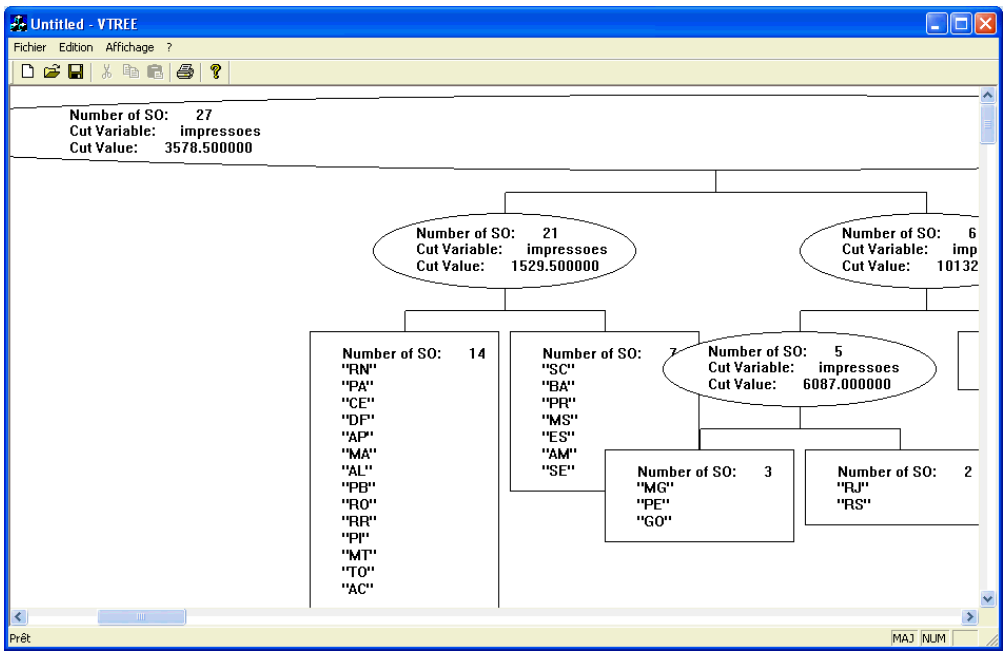


Figura 89 – Método DIV – *output* 2 árvore com *clusters*

Tendo em vista que na etapa de parametrização foi escolhido o *output* extra, pode-se abrir o arquivo "*.sds" gerado no resultado e utilizar o método *View* (Fig. 90) para visualizar o resultado consolidado dos *clusters*, podendo inclusive gerar mais gráficos das informações dos *clusters.*

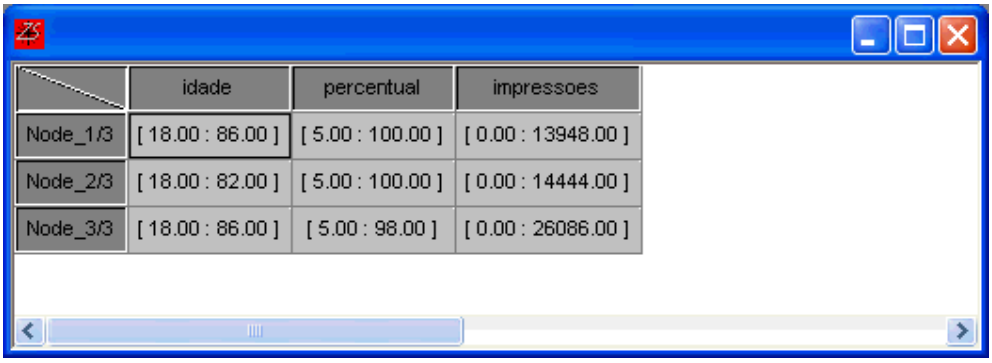


Figura 90 – Método DIV – *output* 3 Visualização (*View*) do arquivo gerado pelo DIV

SCLUST

O primeiro passo para utilizar o SCLUST é arrastar o método para dentro da cadeia, conforme efetuado nos outros exemplos e demonstrado na figura 91.

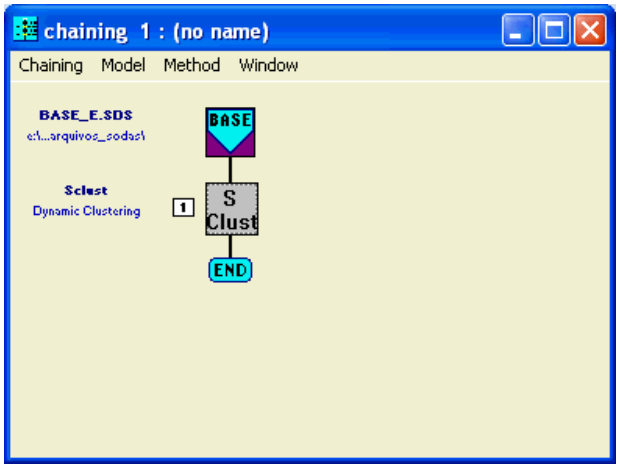


Figura 91 – Método S CLUST

Uma vez com o método incluso na cadeia, deve-se clicar duas vezes em seu ícone para que seja aberta a janela de seleção das variáveis. O SCLUST permite que utilizemos quaisquer tipos de variáveis e as mesmas podem ser utilizadas no mesmo modelo, conforme demonstra a figura 92.

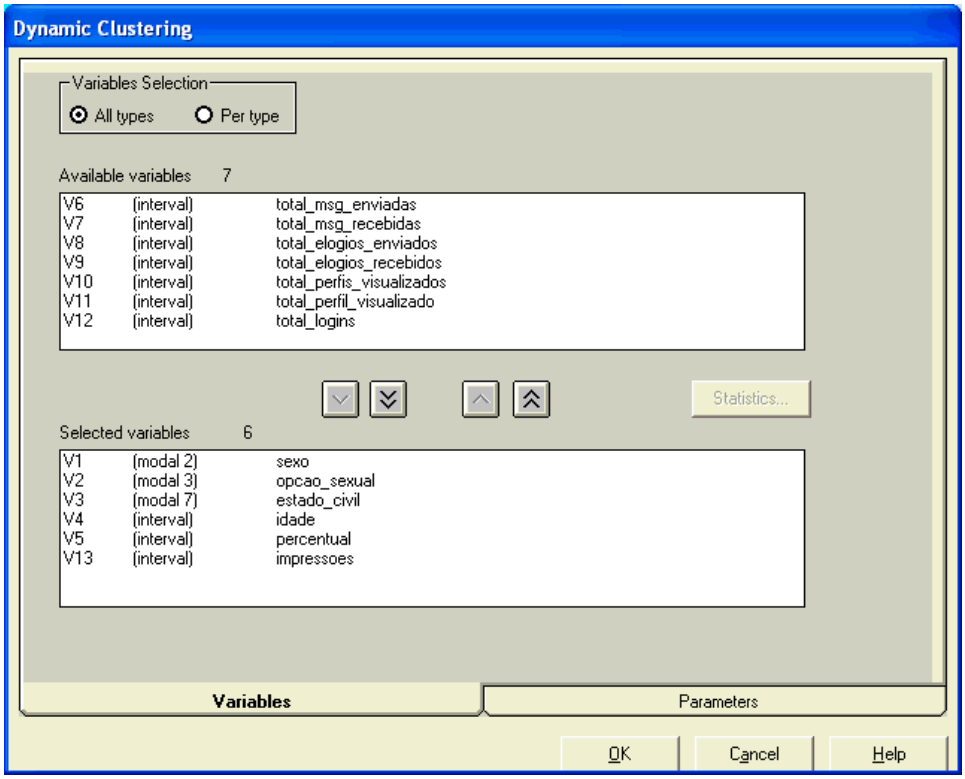


Figura 92 – Método S CLUST Selecionando Variáveis

Uma vez selecionada as variáveis, deve-se acessar a aba "*Parameters*", para configurar o modelo. O SCLUST diferentemente do método DIV exposto anteriormente, oferece uma gama de opções de como trabalhar os dados do modelo, onde é possível ajustar parâmetros importantes, como por exemplo: número de classes, número de interações, tipo de distância utilizada, normalização, entre outros conforme figura 93. Neste exemplo é utilizada a configuração padrão do SCLUST.

O SCLUST igualmente ao DIV também oferece a possibilidade de gerar um novo arquivo "*.sds" com as informações dos resultados, mantendo assim a facilidade na análise dos resultados.

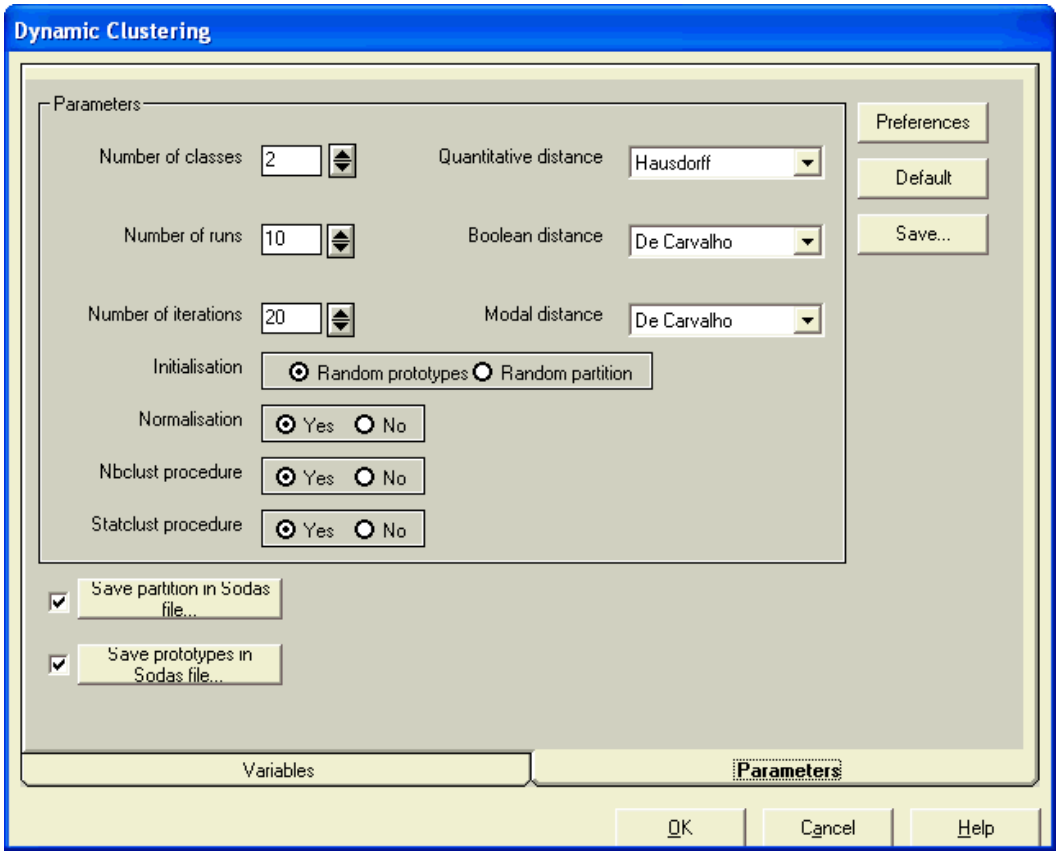


Figura 93 – Método S CLUST ajustando Parâmetros

Depois de efetuada a configuração dos parâmetros, deve-se rodar o método, gerando assim os três *outputs* demonstrados na figura 94, onde o primeiro é um arquivo texto, o segundo são os gráficos e o terceiro são os resultados que estão salvos em outros "*.sds" já configurados no "*View*". É percebido um avanço técnico do módulo SCLUST em comparação com o módulo DIV, uma vez que no módulo anterior o resultado armazenado em "*.sds", somente pode ser visualizado em outra cadeia e o SCLUST já inclui na mesma cadeia o resultado que foi salvo no "*.sds".

No SODAS não é raro encontrar certas diferenças técnicas de módulo para módulo, tendo em vista que por tratar-se de um projeto livre, os módulos não necessariamente são desenvolvidos pela mesma equipe, o que contribui muito para o crescimento e diversidade do *software*.

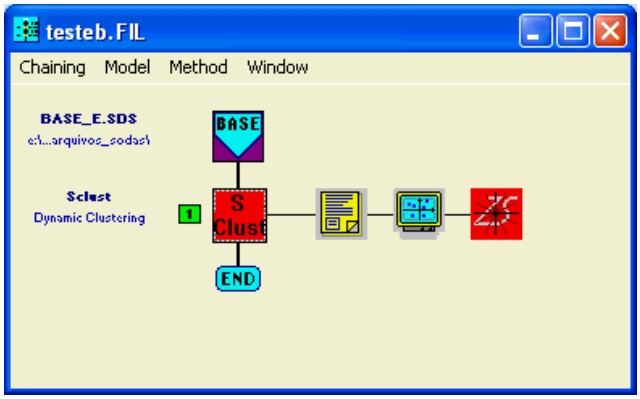


Figura 94 – Método S CLUST após ser rodado

O primeiro *output* gerado exibe um relatório completo de todo o processo efetuado para gerar o resultado, demonstrando informações relevantes, como por exemplo, interações realizadas, permutações, lista de objetos simbólicos, entre outros. (Fig. 95)

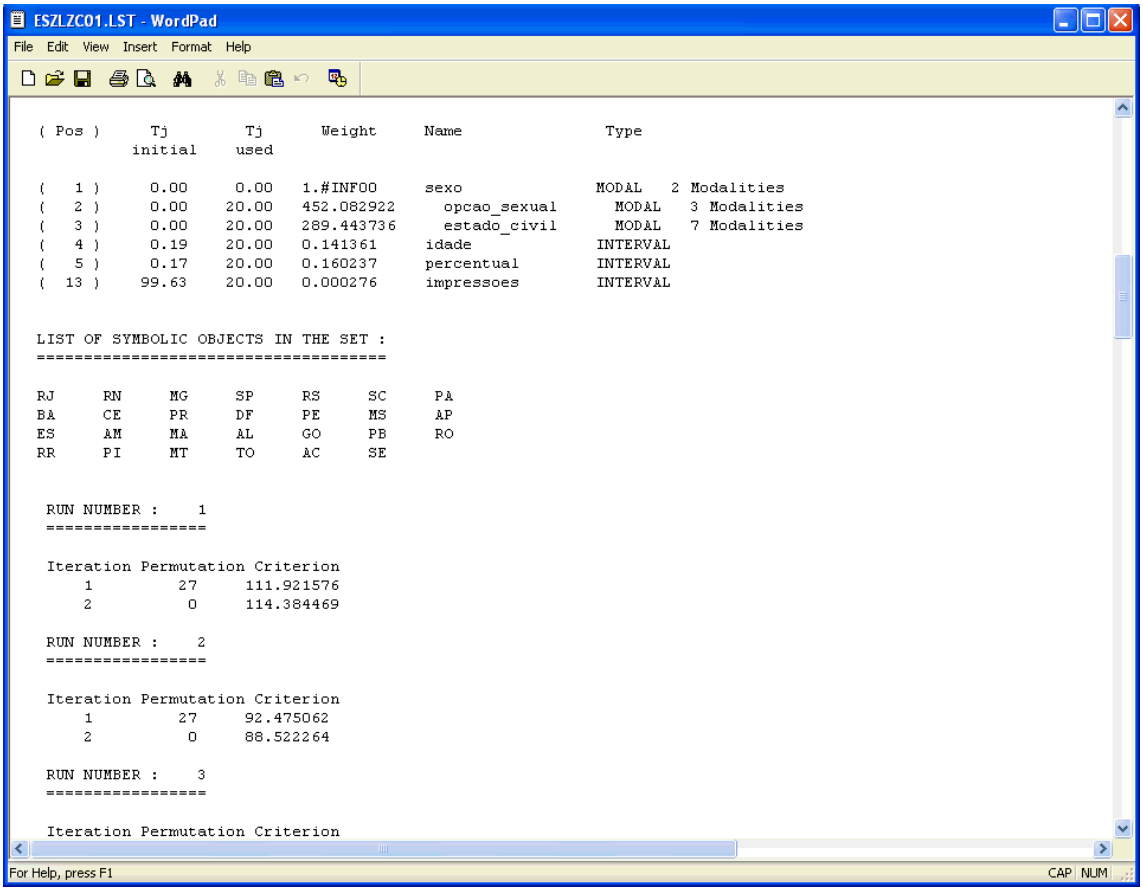


Figura 95 – Método S CLUST – *output* 1 em arquivo texto

Clicando duas vezes sobre o segundo *output*, será exibida uma janela para que o usuário selecione as variáveis que irão compor o eixo horizontal e vertical de seu gráfico. (Fig. 96).

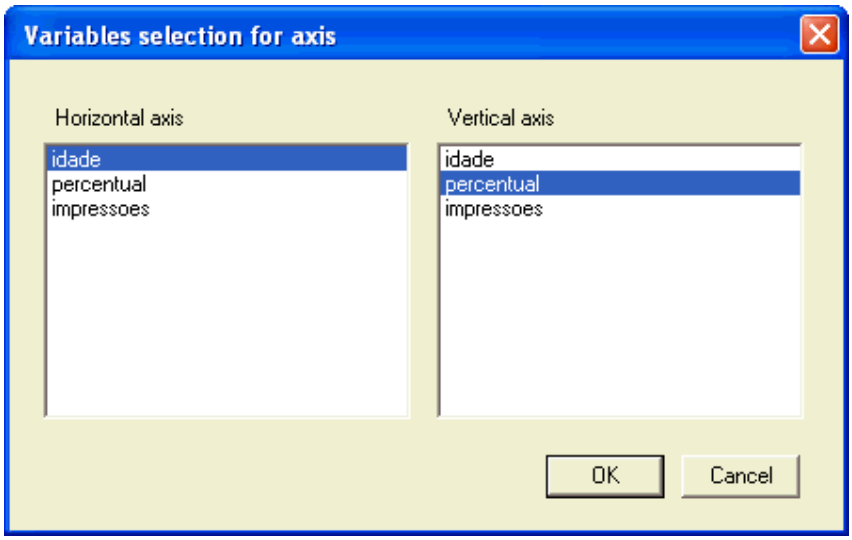


Figura 96 – Método S CLUST – *output* 2 selecionando variáveis para gráfico

Uma vez selecionadas as variáveis e clicando no botão de “Ok”, será exposto o gráfico para o usuário (Fig. 97), com os *clusters* definidos claramente no mesmo.

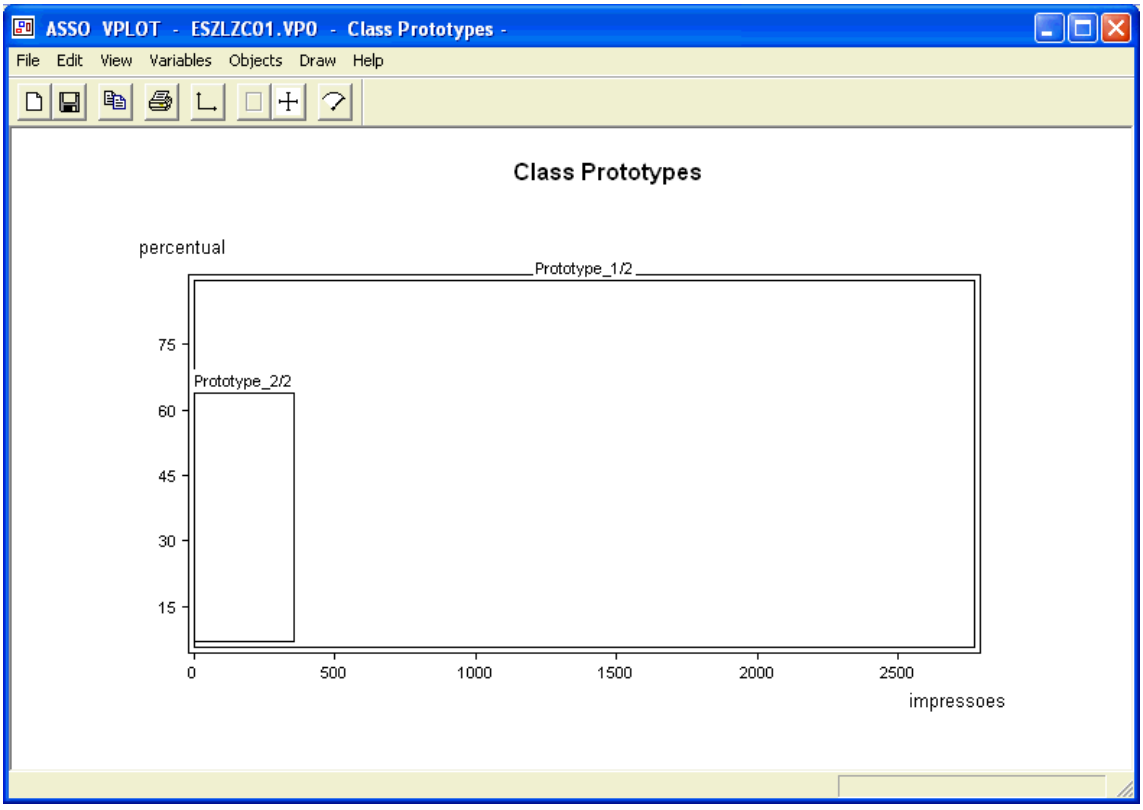
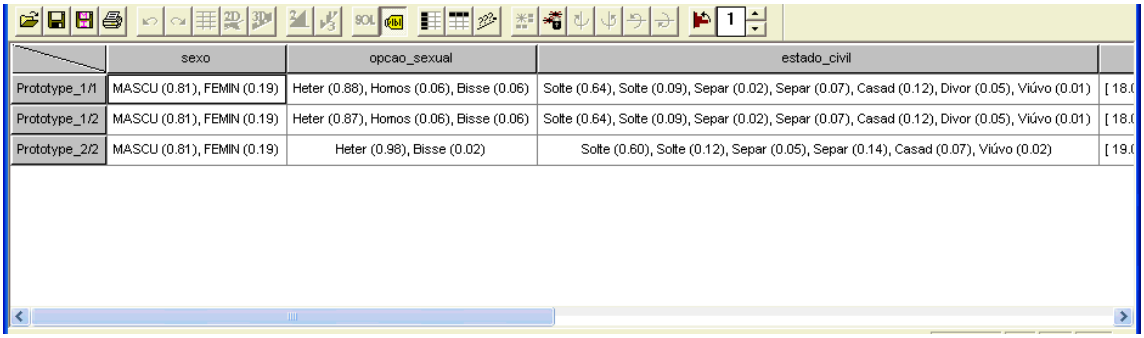


Figura 97 – Método S CLUST – *output* 2 gráfico gerado

O terceiro *output*, gerado pelo método, são as informações dos *clusters* propriamente ditos no método "*View*", onde se pode visualizar as informações de cada *cluster* e também gerar gráficos mais apurados sobre suas composições. (Fig. 98)



	sexo	opcao_sexual	estado_civil	
Prototype_1/1	MASCU (0.81), FEMIN (0.19)	Heter (0.88), Homos (0.06), Bisse (0.06)	Solte (0.64), Solte (0.09), Separ (0.02), Separ (0.07), Casad (0.12), Divor (0.05), Viúvo (0.01)	[18.(
Prototype_1/2	MASCU (0.81), FEMIN (0.19)	Heter (0.87), Homos (0.06), Bisse (0.06)	Solte (0.64), Solte (0.09), Separ (0.02), Separ (0.07), Casad (0.12), Divor (0.05), Viúvo (0.01)	[18.(
Prototype_2/2	MASCU (0.81), FEMIN (0.19)	Heter (0.98), Bisse (0.02)	Solte (0.60), Solte (0.12), Separ (0.05), Separ (0.14), Casad (0.07), Viúvo (0.02)	[19.(

Figura 98 – Método S CLUST – *output* 3 *clusters* importados para ferramenta *View*

SYKSOM

O primeiro passo como demonstrado nos outros exemplo é arrastar o método Syksom para dentro da cadeia, conforme figura 99.

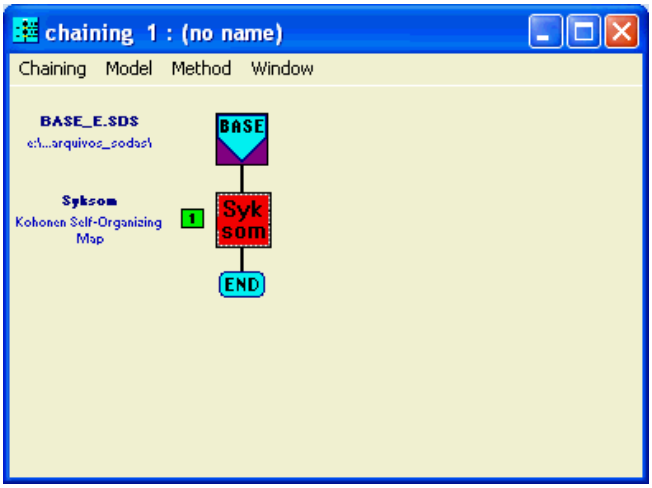


Figura 99 – Método Syksom na cadeia

O segundo passo para a análise é a seleção de variáveis, onde neste método somente devem ser utilizadas variáveis do tipo *Interval.* (Fig. 100)

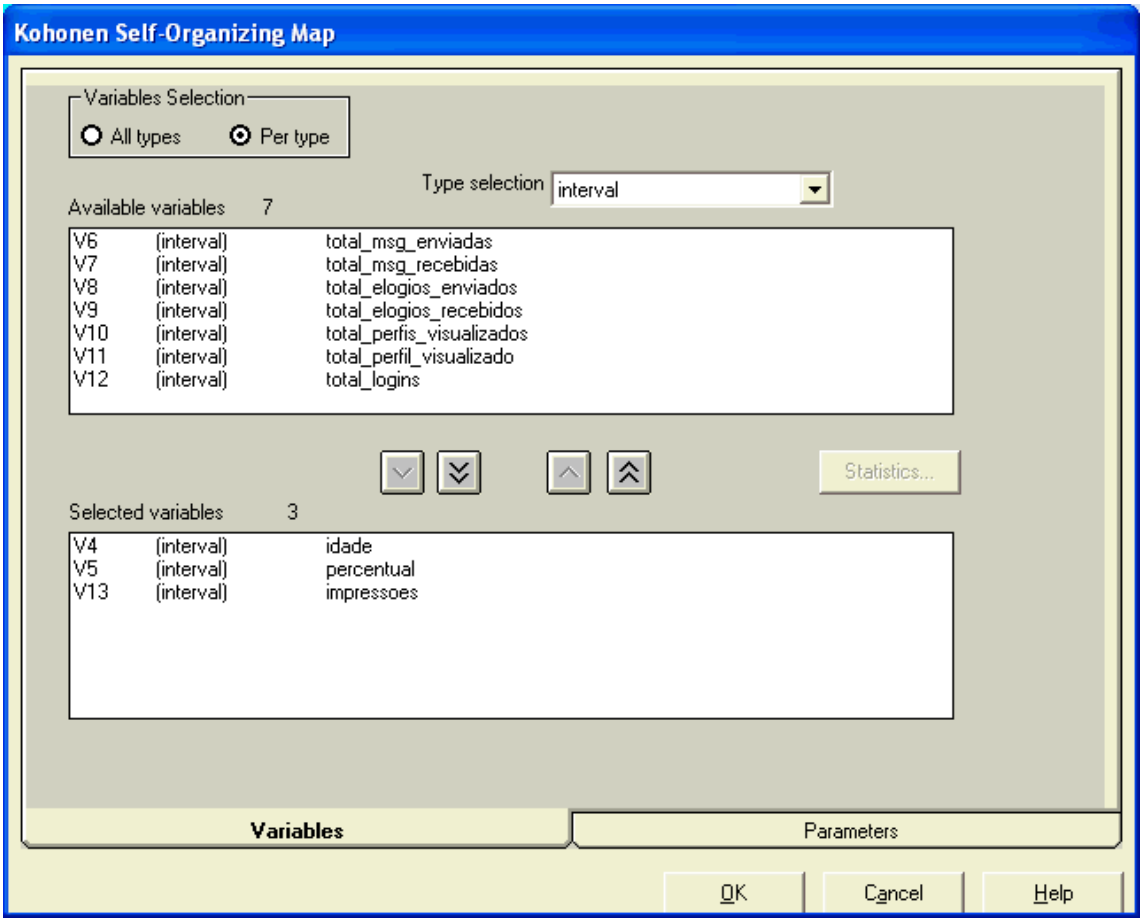


Figura 100 – Método Syksom – Selecionando variáveis

Após selecionar todas as variáveis desejadas, deve-se ajustar a configuração do método através da aba "*Parameters*". Da mesma forma que o SCLUST o Syksom também oferece diversas opções de parametrização do modelo, conforme pode-se verificar na figura 101.

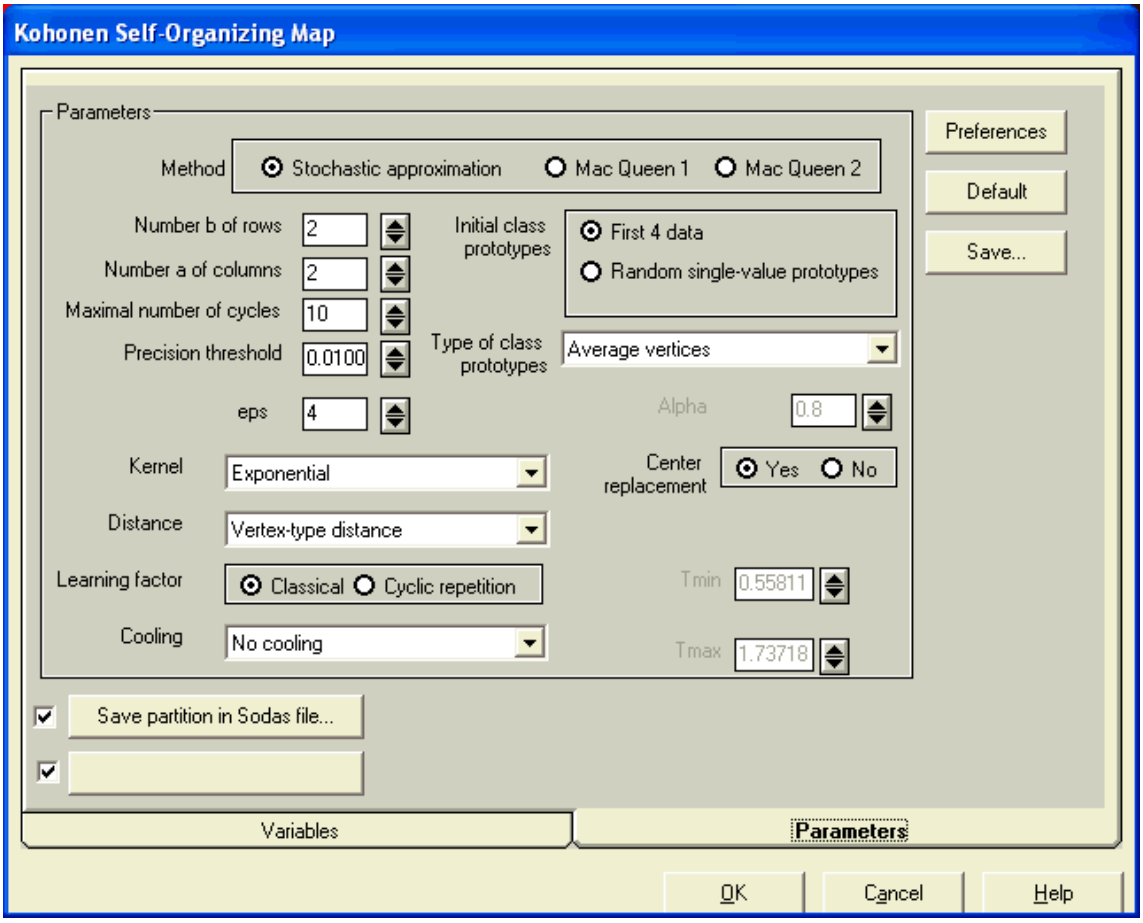


Figura 101 – Método Syksom – ajustando parâmetros

Uma vez configurado o método o próximo passo é rodar o mesmo, gerando os *outputs* para análise, conforme a figura 102. O Syksom, possui 3 *outputs*, sendo 1 deles em forma de texto e 2 em forma gráfica.

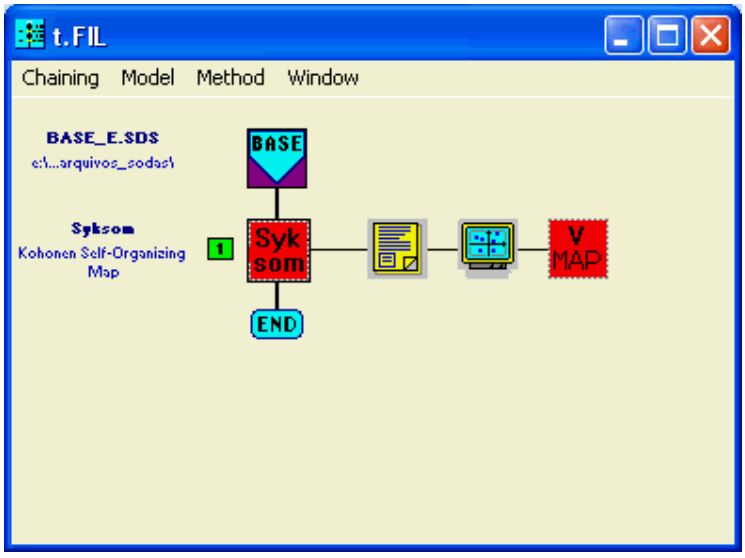


Figura 102 – Método Syksom – *outputs* gerados

O primeiro *output* é o arquivo de texto gerado na execução do método, onde são demonstradas informações dos *clusters*, tais como os valores de máximo e mínimo para cada variável do *cluster*. (Fig. 103)

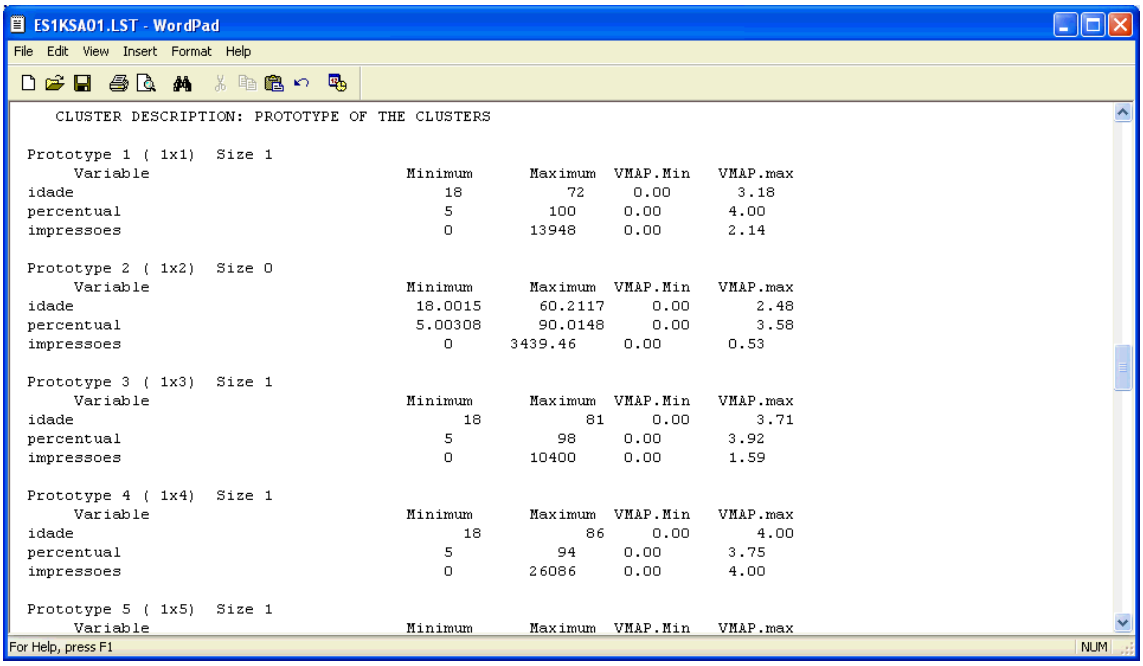


Figura 103 – Método Syksom – *output 1* arquivo de texto

O segundo *output* oferece uma visualização gráfica dos *clusters*, podendo o usuário selecionar o eixo horizontal e vertical do gráfico, conforme figura 104.

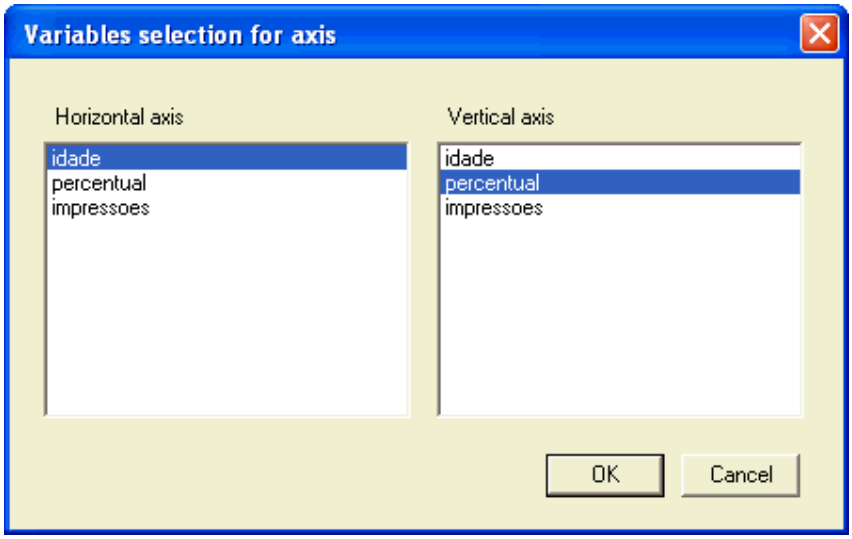


Figura 104 – Método Syksom – *output 2* selecionando variáveis para gráfico

Tendo configurado os eixos, deve-se clicar no botão “ok” e assim será gerado o gráfico representado na figura 105, onde estão definidos os *clusters* de acordo com os eixos escolhidos anteriormente.

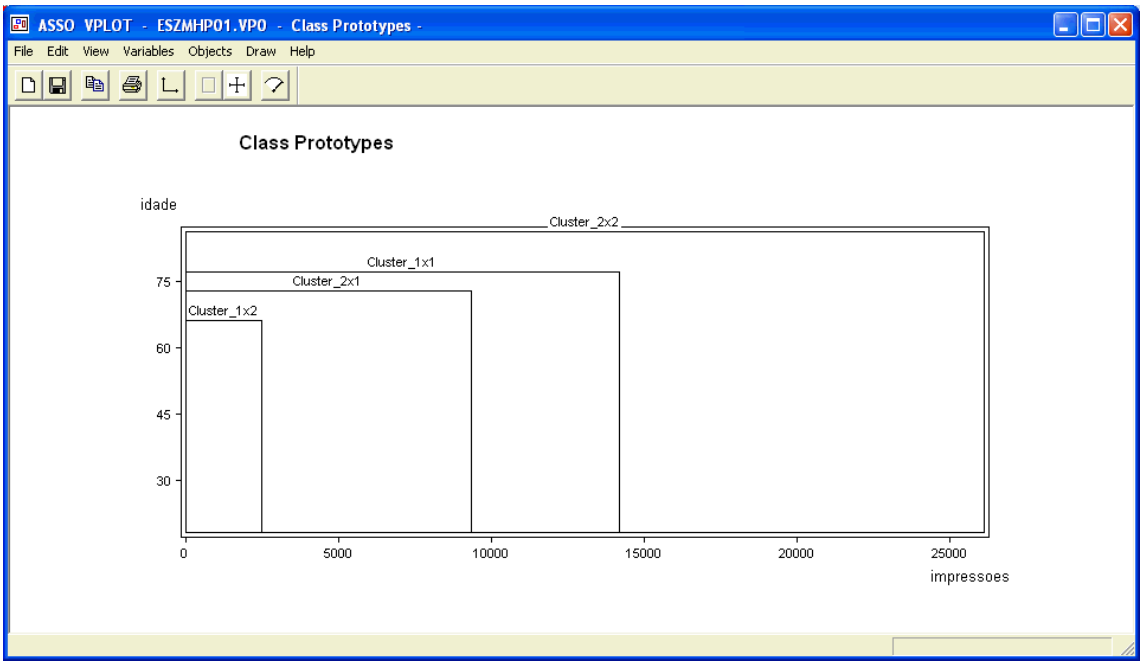


Figura 105 – Método Syksom – *output 2* gráfico

Um dos principais diferenciais gráficos do método Syksom é o terceiro *output* (Fig. 106), onde existe um aplicativo chamado de VMAP, que demonstra os *clusters* gerados. O usuário

ao passar o mouse por cima de cada quadrante do gráfico, são mostradas informações sobre o mesmo, contendo a identificação do neurônio, quantidade de objetos, os números dos objetos e o volume de dados. Pode-se verificar que o aplicativo VMAP do SODAS, ainda não foi traduzido, estando o mesmo em Francês.

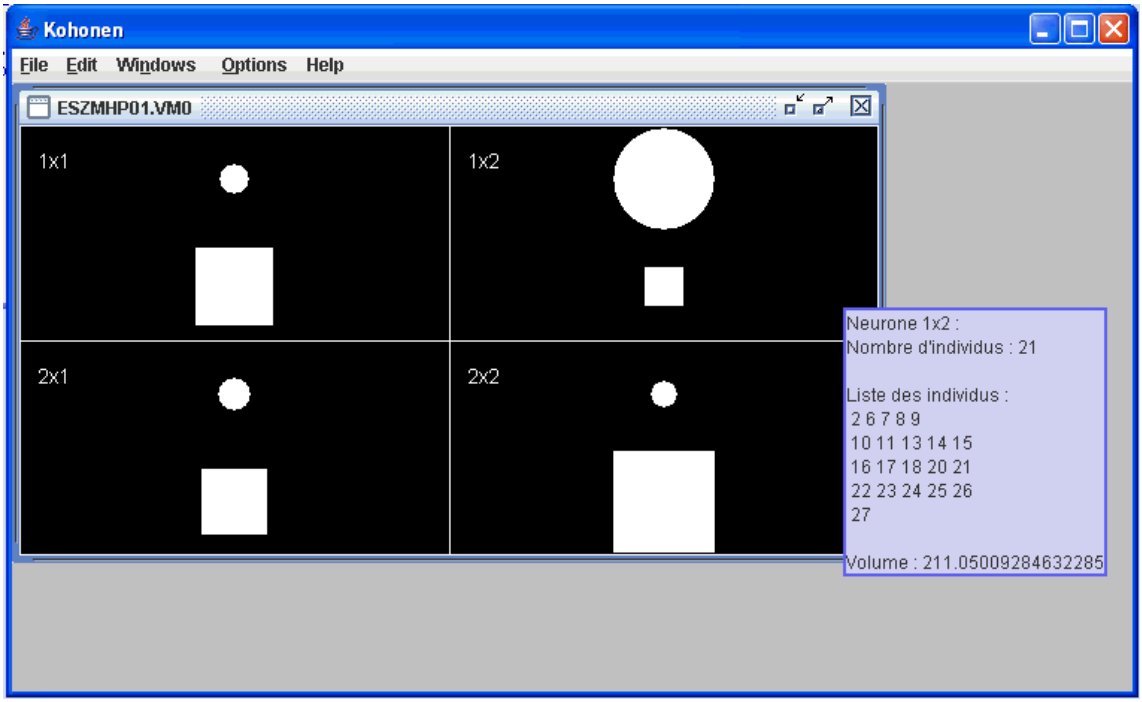


Figura 106 – Método Syksom – *output 3* – gráfico de *clusters*

SCLASS

O método SCLASS é considerado uma "Árvore" de Classificação não supervisionada, onde semelhantemente ao DIV, também trabalha com um método de divisão de *cluster*, onde todos os objetos iniciam em um único *cluster* e ocorrem sucessivas divisões, e cada *cluster* se transforma em dois menores, entretanto o SCLASS apresenta características em seu algoritmo que o diferencia do DIV.

O método SCLASS é um método de *Clustering* em "árvore", onde os nós são divididos recursivamente, baseados em somente uma variável, que é escolhida como ótima pelo modelo. Os cortes são realizados, com a pressuposição de que os dados podem ser modelados conforme o processo não homogêneo de Poisson e estimação através do método Kernel. (SCLASS, 2003)

Para iniciar a utilização do método SCLASS, deve-se primeiramente arrastar o método para cadeia, conforme representado na figura 107.

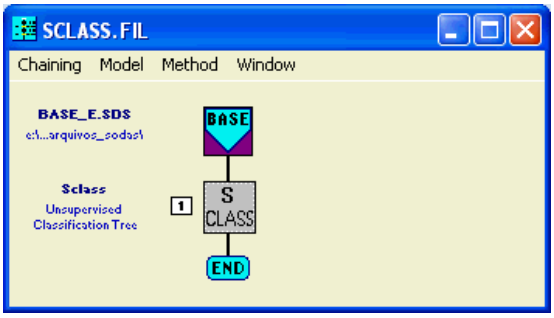


Figura 107 – Método S CLASS na cadeia

O próximo passo é a seleção de variáveis, as quais somente podem ser do tipo *Interval*. Para abrir a janela de seleção deve-se dar um duplo clique no ícone do método (Fig. 108), conforme realizado nos exemplos anteriores.

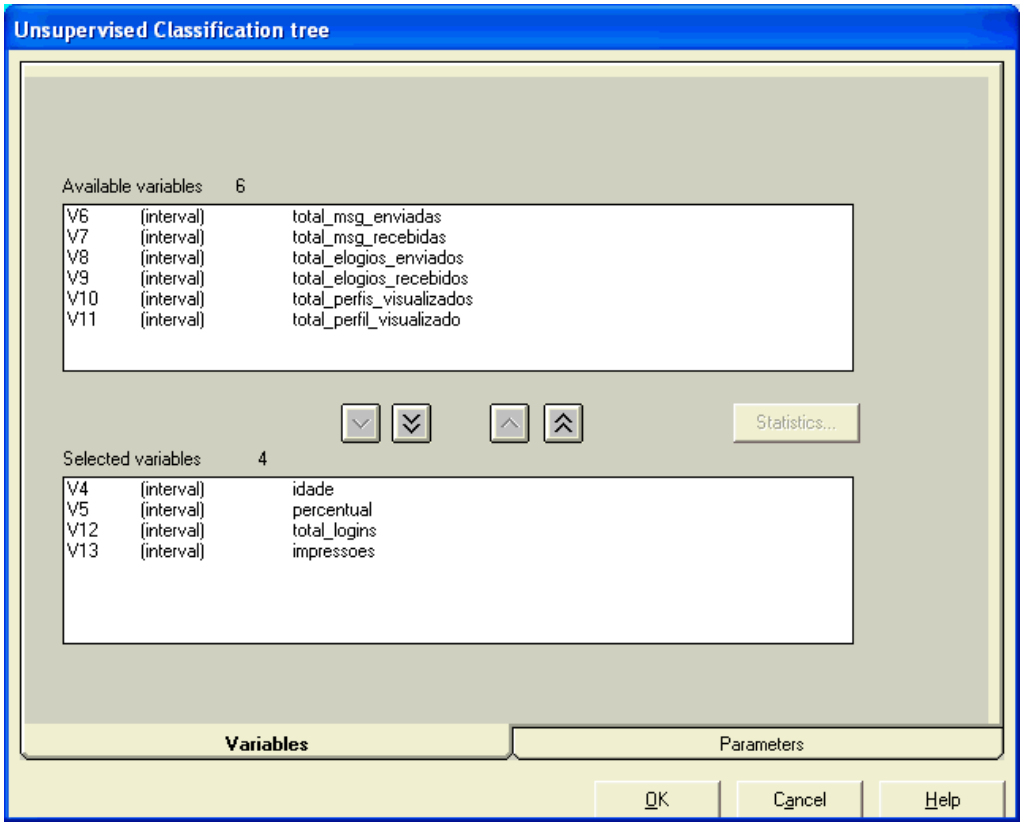


Figura 108 – Método S CLASS – selecionando variáveis

Após efetuada a seleção das variáveis, deve-se acessar a aba "*Parameters"* para efetuar as configurações finais do método, sendo este parâmetro de corte (valor *Alpha)* e o tamanho mínimo de cada nó . (Fig. 109). O método SCLASS também oferece a opção de salvar os resultados gerados em outro arquivo "*.sds", facilitando assim também a análise dos resultados.

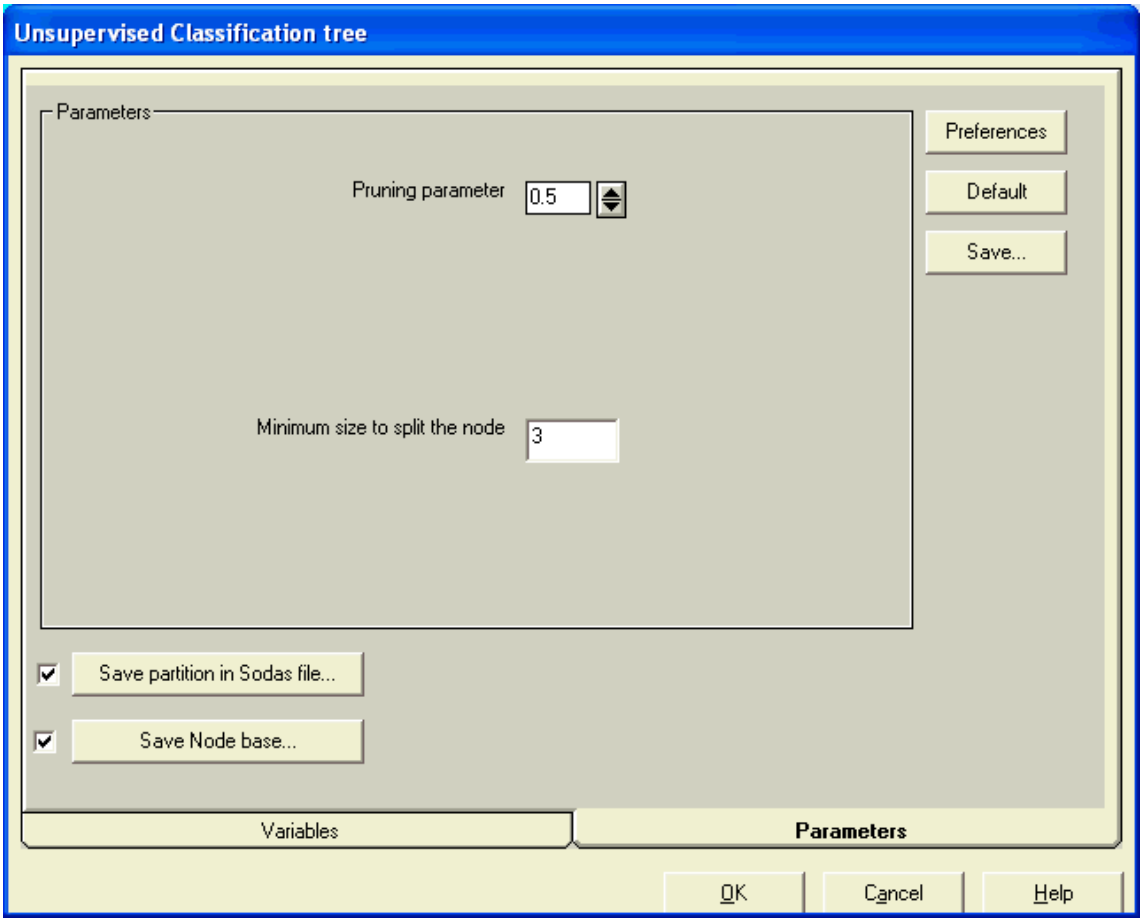


Figura 109 – Método S CLASS – ajustando parâmetros

Depois de efetuar as configurações, deve-se rodar o método, obtendo os *outputs*, demonstrados na figura 110, onde o primeiro será um arquivo de texto, que possui as informações dos *clusters* e as variáveis e parâmetros utilizados para o desenvolvimento do mesmo. O segundo *output* são os resultados consolidados através do *View*, onde se pode exibir e visualizar gráficos dos nós. Já o terceiro *output* exibe a "Árvore" de todos os nós do modelo.

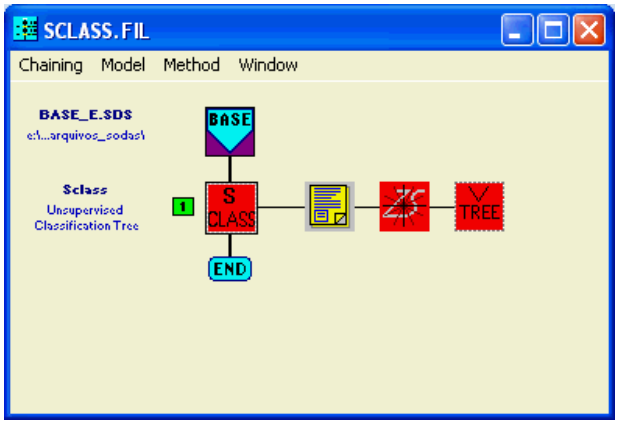


Figura 110 – Método S CLASS – *outputs* gerados

É possível verificar o primeiro *output* através da figura 111, onde se tem como resultados valores como: "*Learning Set*", número de variáveis, mínimo de objetos por nó, valor *alpha*, tão como a relação do nome das variáveis e dos objetos simbólicos da análise.

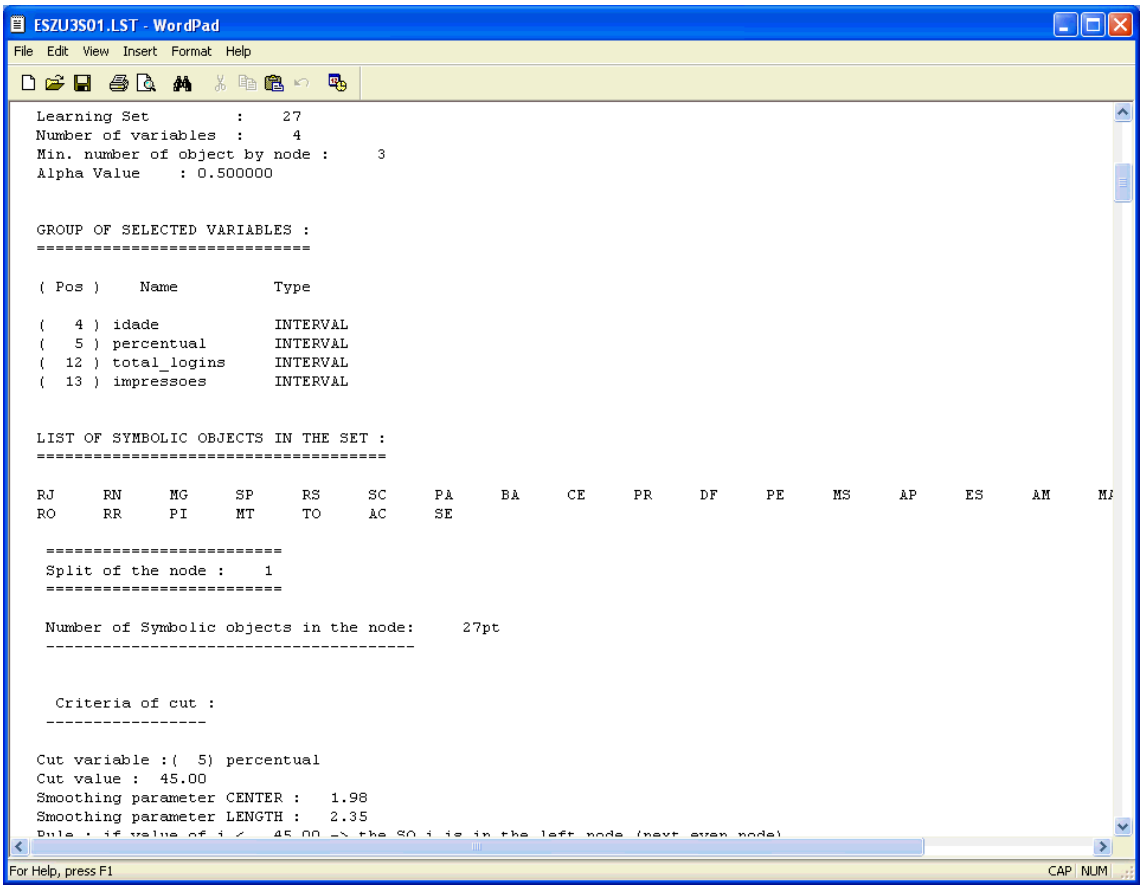
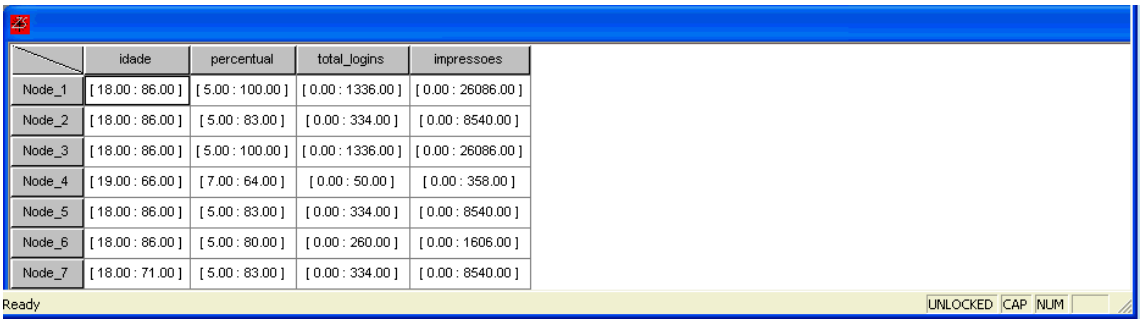


Figura 111 – Método S CLASS – *output 1* arquivo de texto

No segundo *output* representado pela figura 112, pode-se visualizar todos os nós gerados e seus limites para cada tipo de variável, existindo a possibilidade de serem gerados gráficos em 2d e 3d, utilizando as funcionalidades do *View*, já descritas anteriormente nesta dissertação.



	idade	percentual	total_logins	impressoes
Node_1	[18.00 : 86.00]	[5.00 : 100.00]	[0.00 : 1336.00]	[0.00 : 26086.00]
Node_2	[18.00 : 86.00]	[5.00 : 83.00]	[0.00 : 334.00]	[0.00 : 8540.00]
Node_3	[18.00 : 86.00]	[5.00 : 100.00]	[0.00 : 1336.00]	[0.00 : 26086.00]
Node_4	[19.00 : 66.00]	[7.00 : 64.00]	[0.00 : 50.00]	[0.00 : 358.00]
Node_5	[18.00 : 86.00]	[5.00 : 83.00]	[0.00 : 334.00]	[0.00 : 8540.00]
Node_6	[18.00 : 86.00]	[5.00 : 80.00]	[0.00 : 260.00]	[0.00 : 1606.00]
Node_7	[18.00 : 71.00]	[5.00 : 83.00]	[0.00 : 334.00]	[0.00 : 8540.00]

Ready UNLOCKED CAP NUM

Figura 112 – Método S CLASS – *output 2* "nós" visualizados através do método *View*

O terceiro *output* (Fig. 113), exibe a árvore de todos os nós do *cluster*, demonstrando a variável que foi selecionada para "guiar" o modelo (neste caso a variável "percentual"), tão como o seu valor de corte para cada etapa.

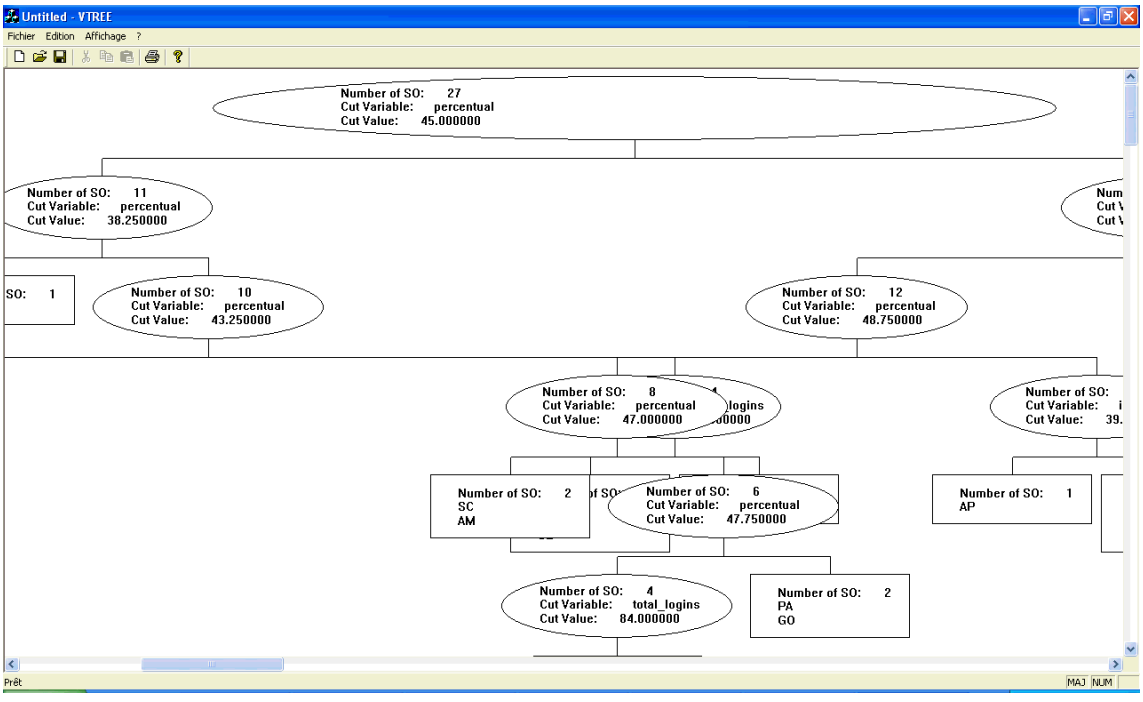


Figura 113 – Método S CLASS – "Árvore" com os nós do cluster

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo