Revista Brasileira de Engenharia Biomédica,
v. 24, n. 2, p. 77-90, agosto 2008

ISSN 1517-3151
http://dx.doi.org/10.4322/rbeb.2012.050

Artigo Original
Recebido em 18/01/2008, aceito em 14/07/2008

MINERSUS – Ambiente computacional para extração de informações para a gestão da saúde pública por meio da mineração dos dados do SUS

MINERSUS – A computational framework for extracting analytical information through data mining of Brazilian public health databases

Ricardo da Silva Santos*
Departamento de Informática em Saúde / UNIFESP
Rua Botucatu, 862 - Ed. Leal Prado, Térreo
04023-062 São Paulo, SP
E-mail: rsantos@compumedica.com.br

Marco Antônio Gutierrez
Serviço de Informática, InCor / FMUSP

*Autor para correspondência

Resumo
O objetivo deste trabalho é apresentar a definição, implementação e validação de um ambiente computacional, denominado MINERSUS, para a produção de informação analítica por meio da mineração das bases de dados dos sistemas de informações do Sistema Único de Saúde (SUS). A partir de uma revisão bibliográfica e de um projeto desenvolvido na Secretaria Estadual de Saúde (SES-SP) (Santos *et al*., 2004 e 2006), levantou-se uma lista de desafios para implantar uma ferramenta analítica na área da saúde. Mediante esta lista foram formuladas premissas que direcionaram a configuração arquitetônica do MINERSUS, e que exigiram a definição e implementação de alguns mecanismos especiais, como o versionamento de tabelas, o controle de mudanças estruturais nos arquivos e os mecanismos específicos incluídos nas atividades de mineração de dados. A utilidade do MINERSUS foi avaliada a partir da operacionalização desse ambiente para responder a questões pertinentes à saúde pública, e a sua usabilidade foi apreciada por meio de uma pesquisa de campo que permitiu a interação do usuário com o ambiente. A avaliação da utilidade confirmou a coerência da informação produzida pelo MINERSUS, comprovando a sua capacidade de extrair informações úteis à gestão da saúde pública. Os resultados da avaliação da usabilidade também foram positivos, comprovando a premissa da facilidade para o usuário realizar a mineração nas bases do SUS. O objetivo principal do trabalho foi alcançado: o ambiente computacional para extração de informação a partir da mineração das bases de dados da saúde pública foi definido, implantado e avaliado.

Palavras-chave: Informática em saúde pública, Planejamento em saúde, Banco de dados, Gerenciamento da informação, Integração de sistemas, Técnicas de apoio à decisão.

Abstract
This paper demonstrates the definition, implementation and validation of a computational framework, called MINERSUS, for the extraction of analytical information through data mining of the Brazilian public health databases. The first point was to define main challenges in order to implement an analytical solution in public health organizations. The process to define main challenges included a bibliographic review and a trial project developed in São Paulo Health Department. These challenges were used to define basic premises which drove the design of MINERSUS, such as special features to table versioning, files structural changes and new mechanism into data mining tasks. MINERSUS was evaluated against two different criteria: utility and usability. We performed the utility evaluation by analyzing the information produced by the framework when answering selected questions from public health management and the usability evaluation has been concluded according to a field survey which evaluated the interaction between users and the framework. The utility evaluation demonstrated the coherency of the information produced by MINERSUS and confirmed its capacity to produce useful information for public health management. The results from the usability evaluation were also positive, and indicated that the framework provides a simple form for users mining data from public health databases. The general objective was achieved: a computational framework to extract information through data mining of public health databases was defined, implemented and evaluated. The contributions are the methodology to build the framework and an implanted framework ready to use.

Keywords: *Public health informatics, Health planning, Databases, Information management, Systems integration, Decision support techniques.*

Extended Abstract

Introduction

There are several successful examples of using computational technologies to produce analytical information at public health organizations; nevertheless, particular challenges in such as organizations hamper the straightforward application of the techniques and tools comprised in those technologies.

The Brazilian National Public Health System is called SUS (Unique Health System). The SUS Information Technology Department, named DATASUS, is responsible for collecting, processing and disseminating information from the SUS and has developed several information systems to achieve this goal. Although these systems record a large amount of information, they are not integrated; each system keeps its own data in an isolated database. Consequently, analytical processing tools cannot be used on these data (Santos et al., 2004 and 2006; Datasus, 2005).

The goal of this work is to define, implement and validate a computational framework for producing analytical information through data mining from SUS databases. This framework is called MINERSUS.

Methods

The first step to define MINERSUS was to establish some premises for defining each component of the framework. The premises derive from the main challenges to implement analytical solutions at the public health and also from specific requisitions from SUS. The diagnosis of these issues was based on bibliographic research and on a project developed at São Paulo Health Secretary (SES-SP) from September 2005 to February 2006 (Santos et al., 2004 and 2006). The MINERSUS architecture shown in Figure 2 was defined considering the established premises. It is composed by a DW (Data Warehouse), an ETL (Extracting, Transforming and Loading) component and an analytical component.

Special requirements for data modeling and a model to handle versioning dimensions were defined for the DW.

The ETL component is shown in Figure 3. It includes several functionalities such as automatically downloading, detection of structural changes on the received files, reading, cleaning, integrating and recording source data in the DW, treating dimension versioning, and log of load process.

The analytical component (Figure 4) is the framework engine responsible for producing information to the public health decision maker. It includes a simple and intuitive interface as well as the ability to support both the OLAP (On-line Analytical Processing) and OLAM (On-line Analytical Mining) technologies, integrated on a clear manner to the user, as part of the analytical process on continuous workflow. This component includes three data mining tasks: clustering, association and classification. Two new methods were defined and implemented on the analytical component: CMDR (Clustering by Minimum Distance Reduction) and ATSV (Association by Time-Space Variation).

The evaluation of the framework is focused into two aspects: utility and usability. The utility evaluation consists on verification that the framework is able to produce analytical information to answer questions proposed by the health manager, and the evaluation of usability consists of confirming the premise that a user, even without advanced statistical knowledge, can easily create analytical reports and data mining models.

Health technicians devised several questions which could be answered through data mining tasks that were implemented on the MINERSUS in order to evaluate the utility. The results produced by the framework were analyzed through an input dataset, which was inserted on an Excel® sheet and adequately manipulated to demonstrate the results.

The usability evaluation was done through field research that evaluated the user interaction with the analytical framework component.

Results and Discussion

The proposed framework was completely implemented, including the analysis and documentation of source databases, the modeling of DW, the implementation of ETL and analytical components, and the load of DW.

The framework allowed the development of reports and data mining models that satisfy all questions elaborated by health technicians. The answers obtained by MINERSUS were assessed by analysis over the input dataset and also observed by a general clinic physician to test its coherency. The coherency on all questions was confirmed.

The assessment of usability confirmed that the analytical component was easy to use. Therefore, 13 persons (87%), without deep statistical knowledge or extensive training, interacted with MINERSUS and produced information to answer specific questions about public health. All volunteers, including those who did not complete all the tasks, considered the process of creating the model and visualizing the result as very easy (53%) or easy (47%).

One of the great contributions offered by MINERSUS is its flexibility to produce reports that contain data from several SUS information systems. The decision maker him/herself, can obtain analytical information in only a few minutes.

The integration of OLAP and OLAM technologies to MINERSUS is much more comprehensive than the simple existence of the two technologies on the same platform; data mining tasks complement the analytical process, giving the health manager more information than those produced by OLAP technology alone.

Conclusion

The general objective was achieved: a computational framework to extract information through data mining of SUS databases was defined, implemented, evaluated and deployed. This study proposes the methodology to build a ready-to-use framework for data mining of SUS databases.

Introdução

A ciência da computação dispõe de uma tecnologia destinada à produção de informação gerencial, denominada OLAP (*On-line Analytical Processing*) (Han e Kamber, 2006), além de uma tecnologia destinada à descoberta de conhecimento, denominada OLAM (*On-line Analytical Mining*) (Han, 1998). Muitos segmentos organizacionais, incluindo a saúde, começam a adotá-las visando obter maior eficiência no planejamento e gerenciamento de suas atividades. No entanto, apesar dos exemplos bem sucedidos na área da saúde, há algumas particularidades nesta área que dificultam a aplicação de ferramentas dotadas de tais tecnologias.

No contexto brasileiro, a saúde pública é implementada pelo Sistema Único de Saúde (SUS), cujos princípios fundamentais são a universalidade do acesso, a integralidade e a igualdade na assistência (Brasil, 1988). A coleta, processamento e disseminação de informações da saúde pública são de responsabilidade do departamento de informática do Ministério da Saúde, denominado DATASUS, que, para cumprir essa tarefa, desenvolveu vários sistemas de informação, tais como Sistema de Informações Ambulatoriais (SIA), Sistema de Informações Hospitalares (SIH), Cadastro Nacional de Estabelecimentos de Saúde (CNES), Sistema de Informações sobre Mortalidade (SIM) e Sistema de Informações de Nascidos Vivos (SINASC). Embora esses sistemas produzam um grande volume de informação, essas informações não estão integradas, uma vez que cada sistema mantém seus dados em bases isoladas, restringindo a obtenção da informação a um único sistema. Neste contexto, as ferramentas sofisticadas de processamento analítico não produzirão o resultado esperado, pois além das bases de dados não serem integradas, também não estão em um formato adequado à produção de informação analítica (Datasus, 2005; Santos *et al*., 2004 e 2006).

Assim, surgiu a hipótese da criação do MINERSUS: um ambiente computacional destinado à produção de informação analítica por meio da mineração das bases de dados do SUS. O objetivo do trabalho é a definição, implementação e validação dos componentes desse ambiente computacional.

Um ambiente para a produção de informação analítica contém uma arquitetura básica como a que está representada na Figura 1. No centro da figura está o armazém de dados, denominado *Data Warehouse* (DW), que armazena os dados em formato específico para a produção de informação analítica (Han e Kamber, 2006; Inmon, 1997). No topo da arquitetura estão as bases de dados que alimentam o DW. O processo de carga dos dados é o componente em que são realizados os procedimentos de limpeza, integração e transformação das diversas bases de dados que alimentarão o *Data Warehouse*. O componente denominado Informações Analíticas corresponde ao mecanismo responsável pela leitura dos dados do DW e pela produção da informação analítica. Finalmente, há um amplo dicionário de dados, denominado Metadados, para auxiliar todos os processos contidos no ambiente de processamento analítico (Inmon, 1997).

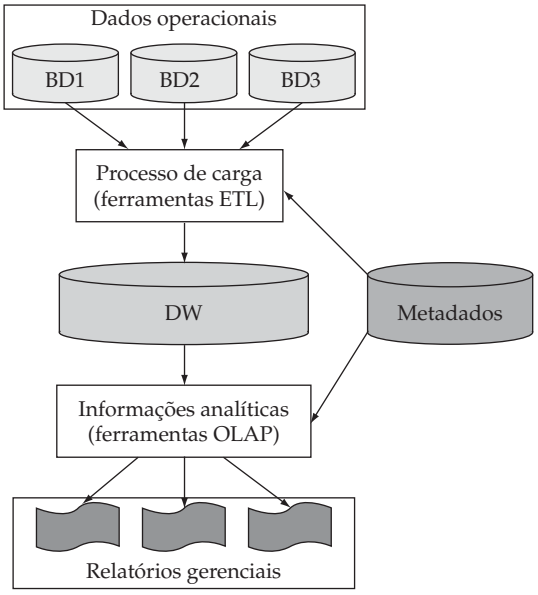

Figura 1. Arquitetura de um ambiente para processamento analítico. ***Figure 1.*** *Analytical processing framework architecture.*

Os dados contidos no DW estão organizados no modelo multidimensional (Han e Kamber, 2006; Kimball, 1998). Este modelo consiste de uma tabela central, denominada tabela fato, e de um conjunto de tabelas periféricas ligadas à tabela fato, denominadas dimensões. Fato é um conjunto de itens de dados, composto por medidas e dados contextuais representando um assunto, uma transação ou um evento do negócio. Dimensões são os aspectos do negócio utilizados para avaliar um fato. Métricas são os atributos numéricos dos fatos que medem o comportamento do negócio em relação às dimensões (Ballard *et al*., 1998).

O processo de carga é o conjunto de atividades realizadas para extração dos dados operacionais e sua inclusão no DW. Nesse processo são realizados procedimentos de limpeza, integração e transformação

dos dados. Existem ferramentas, denominadas ETL (*Extracting, Transforming and Loading*) que auxiliam os procedimentos do processo de carga (Berson e Smith, 1997).

O processo de produção da informação analítica é geralmente realizado por ferramentas baseadas na tecnologia OLAP. Estas ferramentas possuem um conjunto de características específicas, como *Drill-Down*, *Roll-Up*, *Slice-Dice* (Ballard *et al*., 1998; Han e Kamber, 2006).

A tecnologia OLAM, destinada à mineração em bases de dados multidimensionais, também pode ser utilizada para a produção de informação analítica.

O MINERSUS foi projetado para usar ambas as tecnologias na produção da informação, sendo que o componente baseado na tecnologia OLAM contempla três atividades de mineração de dados: classificação, agrupamento e associação.

A classificação é a atividade que determina a qual classe pertence um determinado objeto, dado um conjunto de classes pré-definidas (Goebel e Gruenwald, 1999). Os algoritmos típicos para essa atividade incluem árvore de decisão e redes neurais (Tang e MacLennan, 2005). A atividade de agrupamento consiste em identificar grupos que separam um conjunto de elementos baseados nas características similares destes objetos. Os algoritmos típicos desta atividade incluem o *K-means*, *EM*, *CLARANS*, *BIRCH* e *Minimum Spanning Tree* (Han e Kamber, 2006; Theodoridis e Koutroumbas, 1999).

A associação é a atividade que identifica relacionamentos entre eventos, indicando a existência de regras que determinam um padrão na ocorrência dos eventos. Uma regra de associação com a forma A, B ⇒ C indica que a ocorrência do evento A em conjunto com o evento B, estão associados à ocorrência do evento C. Por exemplo: Pão, Leite ⇒ Manteiga. O principal algoritmo para o estabelecimento de regras de associação é o "*Apriori*" (Han e Kamber, 2006).

Na área da saúde há alguns exemplos bem sucedidos da aplicação da tecnologia OLAP (Berndt e Hevner, 1998; Berndt *et al*., 2003; Breen e Rodrigues, 2001; DeJesus, 1999; Ramick, 2001) e das técnicas de mineração de dados (Bellazzi e Zupan, 2008; Brossette *et al*., 1998; Chae *et al*. 2001; Yang e Hwang, 2006).

Métodos

A arquitetura do MINERSUS (Figura 2) foi desenhada a partir de algumas premissas estabelecidas em função dos principais desafios encontrados na implantação de uma ferramenta analítica para a área da saúde pública. O diagnóstico destes desafios foi baseado em pesquisas bibliográficas e em um projeto piloto desenvolvido na Secretaria de Estado da Saúde de São Paulo (SES-SP) (Santos *et al*., 2004 e 2006).

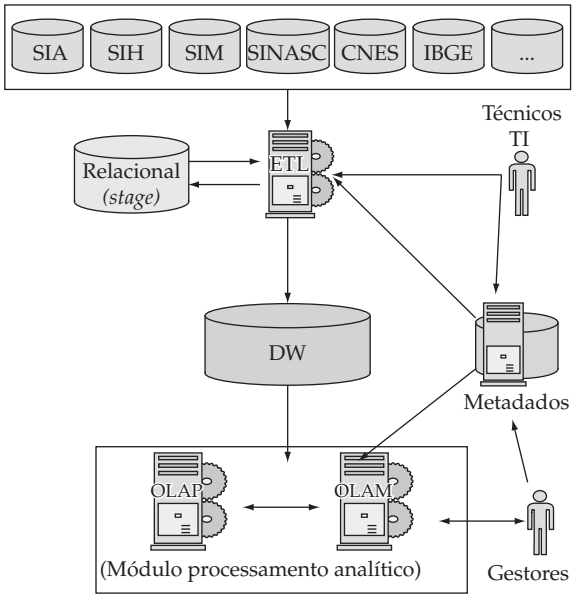

Figura 2. Arquitetura do ambiente computacional.
Figure 2. *Framework architecture.*

Os principais desafios para a implantação de uma ferramenta analítica na área da Saúde estão resumidos na Tabela 1, e as premissas definidas para o MINERSUS estão listadas na Tabela 2.

Data Warehouse - Para atender às premissas adotadas para o MINERSUS não é necessária nenhuma alteração estrutural no processo de modelagem do DW, porém é imprescindível a observância de algumas restrições:

- O esquema multidimensional adotado para o DW é o modelo estrela.
- As métricas dos fatos são atributos numéricos, discretos ou contínuos.
- Existe, para cada fato, pelo menos uma dimensão que representa o tempo e pelo menos uma que representa o espaço (comunidade).
- Existe, para a dimensão que indica o espaço, um atributo normalizador, como população, área, etc.
- Pode existir uma ou mais dimensões versionadas, para as quais devem ser inseridos atributos especiais para o versionamento.

Dimensões versionadas – Algumas tabelas usadas no contexto da saúde pública, como a CID (Classificação Internacional de Doenças) sofrem freqüentes revisões. Se as diferentes versões da tabela não forem tratadas adequadamente, a produção da informação analítica será prejudicada ou até mesmo

Tabela 1. Desafios para implantação de uma solução analítica. ***Table 1.*** *Challenges to implant analytical solution.*

1) Os dados são provenientes de muitas unidades distintas com gestões autônomas, como hospitais, postos de vacinação, secretarias de saúde, etc. Isto apresenta dificuldade e demora na obtenção dos dados.
2) Os dados estão armazenados em uma grande diversidade de formatos.
3) A informação produzida deve ser disseminada para unidades de gestão autônomas, separadas geograficamente e com infra-estrutura computacional bem diferente.
4) Não há soluções comerciais que contemplem os desafios específicos da área.
5) Há, geralmente, limitação de recursos financeiros para investimento em infra-estrutura e soluções sofisticadas.
6) A infra-estrutura atual para a produção da informação analítica está baseada em planilhas do MS-Excel®. É necessário um salto tecnológico muito alto para suportar as sofisticadas ferramentas analíticas.
7) Os dados disponíveis pelo SUS apresentam problemas de integridade referencial, além de valores errados ou sem preenchimento.
9) Falta de documentação para os dados produzidos pelos sistemas de informação do SUS.
10) Existem tabelas, como a CID (Classificação Internacional de Doenças), que sofrem freqüentes revisões, resultando em diferentes versões da mesma tabela.

Tabela 2. Premissas para o ambiente computacional proposto. ***Table 2.*** *Premises to the computer framework proposed.*

1) O MINERSUS produzirá a informação analítica a partir de um DW cujos dados estão organizados no formato multidimensional.
2) O MINERSUS deve prover um mecanismo para efetuar a carga dos dados, dotado de funcionalidades destinadas à solução dos problemas de qualidade dos dados.
3) O MINERSUS deve prover um mecanismo que facilite a análise tempo-espacial de eventos, como por exemplo, epidemias.
4) O MINERSUS deve prover um mecanismo para tratar adequadamente o versionamento das tabelas.
5) O MINERSUS deve integrar os recursos OLAM aos recursos OLAP de maneira transparente ao usuário, como parte do processo analítico num fluxo gradativo e contínuo.
6) O MINERSUS deve prover uma forma simples para o usuário final criar, alterar e processar os modelos de mineração de dados, em tempo real, como parte do processo de análise, sem a necessidade de um especialista para preparar os dados ou configurar os parâmetros dos modelos de mineração.

inviabilizada. Há um exemplo que representa esta situação: uma revisão na tabela de procedimentos ambulatoriais em 10/1999. A interface do sistema para consultar os procedimentos realizados possui duas entradas para o campo "procedimentos": uma que permite a consulta dos procedimentos realizados antes de 10/1999 e outra para consultar os procedimentos realizados após esse período. Devido à existência de duas versões da tabela de procedimentos é impossível realizar uma análise que considere todo o ano de 1999.

Embora o problema de versionamento de dimensões já tenha sido abordado em alguns trabalhos (Ballard *et al*., 1998; Eder e Koncilia, 2001), para o MINERSUS foi proposto um modelo alternativo, focado na simplicidade da implementação.

O modelo proposto consiste, basicamente, na inclusão de atributos especiais na dimensão versionada. Cada dimensão versionada é descrita da seguinte forma:

$$Dv = \{ID, DK, UA_1, UA_2, ...UA_N, VI, JK, JW\} \quad (1)$$

onde:

- ID é o elemento de ligação da dimensão com o fato, formado pelo conteúdo do DK acrescido da data de mudança de versão;
- DK é a chave do registro para o usuário, em que apenas uma linha estará válida em um determinado momento;
- UA são aos atributos definidos pelo usuário;
- VI é o atributo que indica se um determinado registro está ativo ou não;
- JW determina, nos casos de divisão ou agrupamento de registros, o percentual de participação de um registro derivado em um registro agrupado. Por exemplo, numa tabela onde um registro é dividido em 4 novos registros, pode ser atribuído o valor 0,25 para o JW de cada novo registro;
- JK armazena o valor da chave dimensional (DK) de um registro relacionado. Por exemplo, se um determinado registro foi dividido em 4 novos tipos, esses novos tipos recebem em JK o valor de DK do registro antigo.

O componente ETL – A arquitetura geral definida para o componente ETL do MINERSUS é apresentada na Figura 3.

O gerenciador de *download* consiste num mecanismo para efetuar automaticamente o *download* dos arquivos disponibilizados pelo DATASUS. Esse mecanismo contém uma função destinada à análise de volume, para evitar a transferência de arquivos corrompidos, e uma agenda que permite a programação de um dia e horário para a transferência dos arquivos.

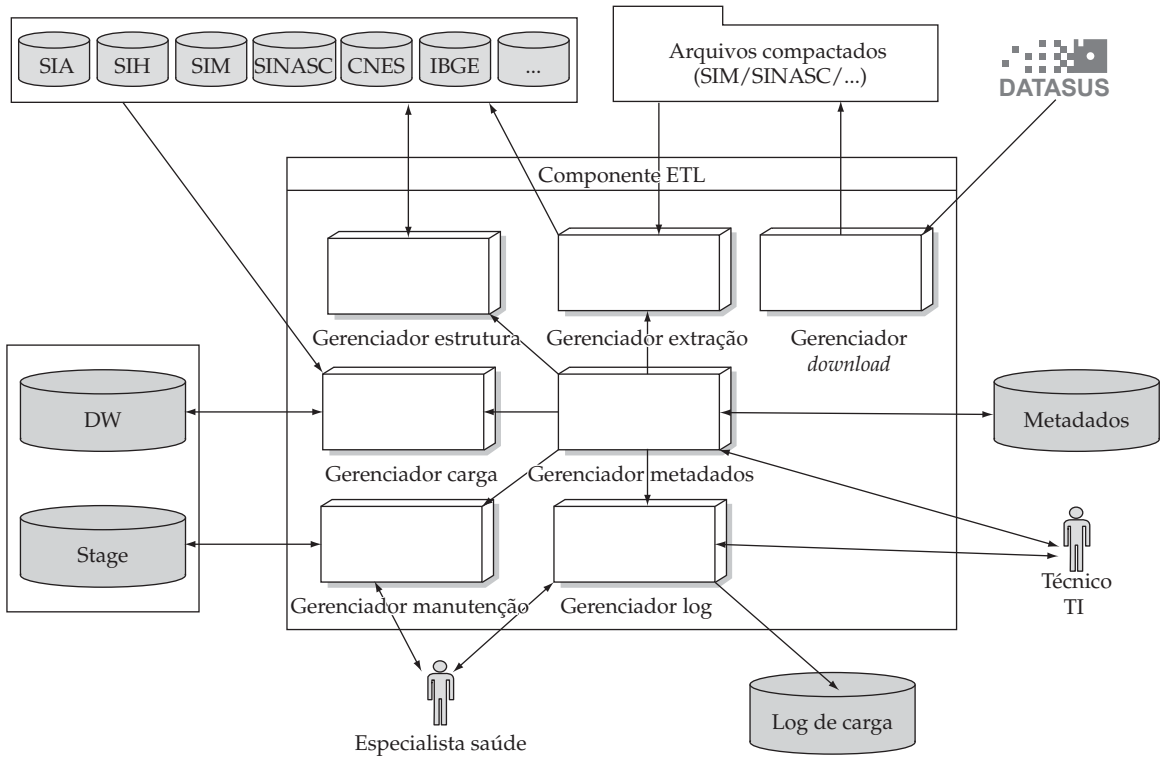

Figura 3. Arquitetura do componente ETL. ***Figure 3.*** *ETL component architecture.*

O gerenciador de extração é responsável pela extração dos arquivos compactados.

O gerenciador de estrutura é um mecanismo para detectar alterações na estrutura dos arquivos recebidos. Este mecanismo sinaliza as diferenças estruturais e oferece algumas funcionalidades para o devido ajuste da estrutura, impedindo a carga dos dados se a estrutura não estiver correta.

O gerenciador de carga é o mecanismo responsável pela execução dos mapeamentos, ou seja, ele efetua a leitura dos dados de origem, o tratamento, a consolidação e a gravação dos dados no DW. Nesse mecanismo estão inseridas as tarefas de limpeza, padronização, consistência, integração, transformação dos dados e análise do versionamento.

O gerenciador de metadados é o módulo responsável pelo cadastramento das configurações utilizadas pelos diversos módulos do componente ETL incluindo as regras de integridade, as dimensões que são versionadas, as fórmulas de transformação e o mapeamento dos objetos.

O gerenciador de *log* tem como objetivo manter o registro e permitir a visualização de todas as ocorrências no processo de carga.

O gerenciador de manutenção corresponde a uma interface para manutenção de objetos do DW. A alteração de objetos no DW não é uma situação comum, porém isto ocorre no contexto da saúde pública pela existência de dimensões cujo conteúdo não está armazenado em um arquivo físico, mas publicado em documentos oficiais impressos.

O componente analítico – O componente analítico é o elemento do MINERSUS responsável pela produção da informação analítica para o gestor da saúde. De acordo com as premissas estabelecidas, o componente analítico deve contemplar uma interface simples e intuitiva, além de suportar as duas tecnologias, OLAP e OLAM, integradas de forma transparente ao usuário, como parte do processo analítico num fluxo gradativo e contínuo. A arquitetura proposta para esse componente está representada na Figura 4.

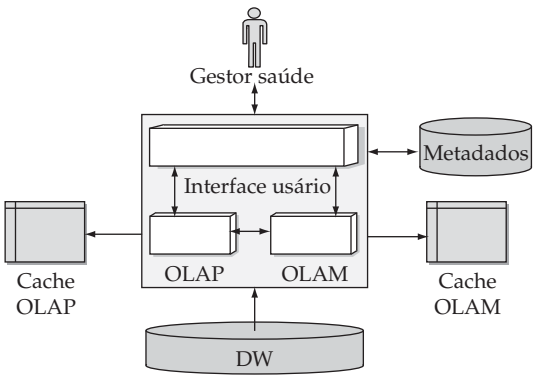

Figura 4. Arquitetura do componente analítico. ***Figure 4.*** *Analytical component architecture.*

A interface do usuário – A interface do usuário é o módulo responsável pela configuração e apresentação dos resultados dos relatórios OLAP e dos modelos de mineração de dados (Figura 5). Há um único módulo destinado à configuração dos modelos, que pode ser um relatório OLAP ou um modelo de mineração de dados. Para exibição dos resultados, devido às características específicas de cada tecnologia, existem dois módulos distintos, um para OLAP e outro para OLAM.

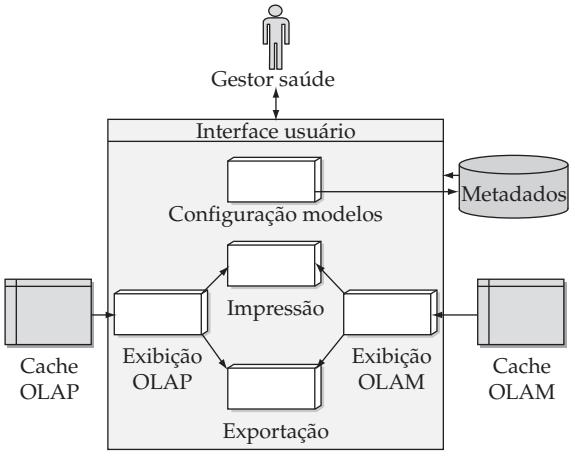

Figura 5. Arquitetura do componente interface do usuário. ***Figure 5.*** *User interface architecture.*

A integração entre as tecnologias OLAP e OLAM vem sendo estudada há uma década (Han, 1998; Han *et al*., 1997). No MINERSUS, a chave para esta integração consiste na representação conceitual do conjunto de dados para a produção da informação analítica. Esta representação conceitual está fundamentada no seguinte princípio: o conjunto de dados de entrada para a produção de informação analítica é obrigatoriamente um cubo de dados derivado de um ou mais cubos contidos no DW. Este princípio conduz a duas deduções lógicas:

- Se um relatório OLAP é a apresentação dos dados num formato de cubo, logo, um relatório OLAP pode ser submetido como entrada para um modelo de mineração de dados;
- Se a entrada para um modelo de mineração de dados é um cubo, logo, os dados derivados de um modelo de mineração podem ser apresentados como um relatório OLAP.

De acordo com estas deduções, a integração OLAP-OLAM é implementada pela interface do usuário por meio de uma funcionalidade que permita o acionamento direto de uma atividade de mineração de dados a partir de um relatório exibido em tela, assim como a exibição de um relatório OLAP pode ser acionada a partir de um modelo de mineração.

Os módulos de exibição, OLAP e OLAM, são os responsáveis pela apresentação dos resultados que estão armazenados em uma área denominada *cache*, implementada por uma tabela de banco de dados. Esses módulos não se restringem apenas à apresentação dos resultados, eles suportam as operações inerentes às ferramentas OLAP e OLAM, tais como *Drill-Down*, *Roll-Up*, *Slice-Dice*, filtros dinâmicos, visualização em gráfico, etc.

Os módulos para processamento dos modelos – Os módulos OLAP e OLAM são os responsáveis pelo processamento dos modelos configurados e pelo armazenamento dos resultados na tabela *cache* correspondente.

O processamento de um modelo de relatório pelo módulo OLAP consiste em transformar o modelo configurado pelo usuário em um comando da linguagem SQL[1] (*Structured Query Language*), e em submeter esse comando ao sistema gerenciador de banco de dados e armazenar o resultado no *cache*.

A tarefa do módulo OLAM consiste em traduzir o modelo configurado pelo usuário para o formato especificado pela API (Interface de Programação de Aplicativos)[2] correspondente, acionar a execução desse modelo e armazenar o resultado na área de *cache* para posterior exibição. O módulo utiliza duas API para execução dos modelos: a OLEDB-DM[3] e uma API proprietária desenvolvida especificamente para o MINERSUS. A arquitetura proposta para este módulo está representada na Figura 6.

A atividade de agrupamento é implementada por meio da API OLEDB-DM, usando os algoritmos *K-means* e *EM*, e pela API proprietária que implementa mais duas funcionalidades definidas para atender as premissas estabelecidas para o MINERSUS: um algoritmo opcional para agrupamento, denominado ARMD (Agrupamento pela Redução das Menores Distâncias), e um mecanismo para análise de distorções dos grupos criados.

[1] Linguagem padrão para consulta e manipulação de dados implementada pelos sistemas gerenciadores de banco de dados.

[2] API é um conjunto de rotinas e padrões estabelecidos por um *software* para utilização de suas funcionalidades por programas aplicativos que não querem envolver-se em detalhes da implementação do *software*, mas apenas usar seus serviços.

[3] OLEDB-DM (*Object Linking and Embedding Database for Data Mining*) é uma API projetada pela Microsoft®, que implementa conceitos comuns de mineração de dados. A filosofia-chave da OLEDB-DM é mapear conceitos do mundo de banco de dados relacional para o mundo da mineração de dados (Tang e MacLennan, 2005).

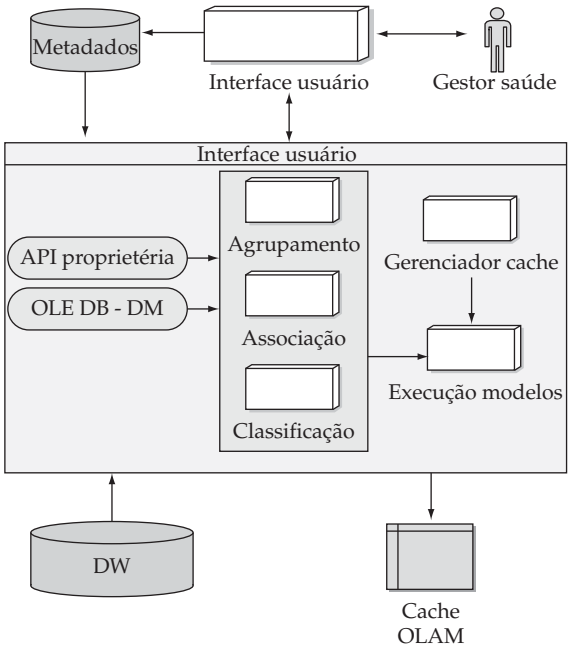

Figura 6. Arquitetura do módulo OLAM. ***Figure 6.*** *OLAM engine architecture.*

O ARMD é adequado para trabalhar com valores dissonantes, uma situação comum nos dados da saúde pública, que pode prejudicar a coerência dos grupos criados. Por exemplo, se deseja criar 3 grupos a partir do conjunto de valores V = {400, 450, 480, 490, 780, 781, 782, 783, 900, 1.845, 2.450, 10.720}. O algoritmo deve isolar o valor discrepante, formando os grupos: V_1 = {400, 450, 480, 490, 780, 781, 782, 783, 900}, V_2 = {1.845, 2.450} e V_3 = {10.720}. O ARMD está fundamentado no *Minimum Spanning Tree* (Theodoridis e Koutroumbas, 1999) e pode ser descrito, de maneira geral, pelo seguinte algoritmo:

1) Ordenar as ocorrências da dimensão de acordo com a sua métrica correspondente;
2) Criar um vetor contendo os valores distintos e ordenados da métrica;
3) Criar *K* faixas a partir do vetor;
4) Para cada ocorrência da dimensão;
5) Para *i* = 1 até *K*;
6) Se o valor da métrica ≤ limite superior da faixa (*i*);
7) Adicionar a ocorrência da dimensão ao Grupo (*i*);
8) Sair do Loop;
9) Fim Se;
10) Fim Para;
11) Fim Para.

Onde *K* é o número de grupos desejados e cada faixa é constituída por um valor que representa o limite inferior e outro que representa o limite superior.

O mecanismo para detecção de distorções nos grupos formados considera como distorção elementos com métricas distantes da média geral, períodos cuja média das métricas de seus elementos esteja distante da média geral, elementos que mudam de grupos ao longo do tempo e grupos que apresentam concentração de elementos.

A atividade de associação é implementada por meio da API OLEDB-DM, usando o algoritmo *Apriori*, e pela API proprietária que implementa um mecanismo opcional para estabelecer associações baseadas na análise tempo-espacial, denominado AVTE (Associação pela Variação no Tempo-Espaço).

O princípio básico da AVTE está fundamentado no seguinte axioma: "dois elementos estão correlacionados se os seus padrões de comportamento são semelhantes no tempo e no espaço". Considerando um cubo de dados constituído por N dimensões e N métricas, e contendo pelo menos uma dimensão que representa o tempo ou o espaço, dois elementos estão correlacionados se, e somente se, a similaridade entre as taxas de variações similares ao longo das unidades da dimensão tempo ou espaço for maior que um limite estabelecido como grau de confiança. Formalmente, pode ser definido como:

$$A \Rightarrow B \Leftrightarrow |S_T(A, B)| \geq C \vee |S_E(A, B)| \geq C \qquad (2)$$

onde: S_T = Similaridade em função do tempo, S_E = Similaridade em função do espaço e C = Grau de confiança, que compreende uma faixa de valores de 0 a 1.

O processo de produção de regras de associação por meio da AVTE é dividido em duas etapas: o cálculo da variação e a produção das regras.

O cálculo da variação consiste em estimar as variações de tempo e espaço que cada elemento apresenta e, conseqüentemente, em eliminar aqueles cuja variação seja inferior a um limite estabelecido como parâmetro, denominado suporte. A primeira tarefa para calcular a variação de um elemento é definir as faixas de variações e atribuir um símbolo específico para cada faixa. Se considerarmos 20 faixas de variação, o fator de variação será de 5% e as faixas serão representadas pelo conjunto {A, B,..., T}, que representam as variações: A = {0 a 5%}, B = {5,01 a 10%} e T > {90%}. Ao atribuir, para cada unidade de tempo ou espaço, o símbolo referente à sua variação, será formada uma *string* de símbolos denominada *string* de variação. O cálculo da variação equivale à diferença do valor 100 pelo percentual de ocorrência do maior símbolo.

A produção das regras de associação consiste em calcular a similaridade entre as *strings* de variações de dois elementos, que é obtida pela média das distâncias dos símbolos, ou seja, quanto um símbolo de uma

string X está distante do seu correspondente na *string* Y. Serão descartados todos os pares de candidatos cuja similaridade entre as *strings* seja inferior ao limite estabelecido como grau de confiança.

A atividade de classificação é implementada pela técnica de árvore de decisão, por meio da OLEDB-DM. A atividade de classificação implementada no MINERSUS, apesar do nome, não realiza a predição de um caso em particular, mas encerra-se na criação da árvore de decisão, pois para o gestor da saúde pública o importante é compreender o perfil de um determinado evento tendo em vista o direcionamento dos recursos ou a elaboração de programas de saúde.

Avaliação do MINERSUS

Alguns estudos destinados à avaliação de ferramentas ou técnicas de mineração de dados foram analisados para subsidiar a avaliação do MINERSUS. Muitos destes estudos concentram-se apenas na avaliação das técnicas de mineração (Mcgarry, 2005; Soares *et al*., 2000), enquanto outros se destinam à avaliação de ferramentas (Collier *et al*., 1999; Goebel e Gruenwald, 1999). A avaliação do MINERSUS está focada nos critérios de avaliação de ferramentas de mineração de dados, especificamente em duas, das quatro categorias de avaliação propostas por Collier *et al*. (1999): utilidade e usabilidade. A avaliação da utilidade consiste em verificar se o MINERSUS consegue integrar as bases de dados e produzir informações capazes de responder a questões pertinentes à gestão da saúde pública. A avaliação da usabilidade consiste em confirmar a premissa de que um usuário, mesmo sem conhecimentos avançados em estatística, pode criar, com facilidade, relatórios analíticos e modelos de mineração de dados. As demais categorias de avaliação propostas por Collier *et al*. (1999) (desempenho e atividades auxiliares) foram avaliadas empiricamente, tanto na implementação dos componentes do MINERSUS, quanto nas avaliações de utilidade e usabilidade. O desempenho foi considerado satisfatório pelos desenvolvedores do *software*, pelos voluntários da pesquisa de usabilidade e pelos técnicos da saúde envolvidos na avaliação.

Foram adotadas duas práticas para avaliar a utilidade do componente OLAP do ambiente. A primeira consistiu no desenvolvimento de relatórios capazes de produzir a informação contida nas planilhas desenvolvidas pelos técnicos da SES-SP no ano de 2005, para suprir a demanda de informações do gestor da saúde. A segunda prática foi solicitar aos técnicos da saúde da SES-SP a elaboração de algumas questões para serem respondidas por meio do componente OLAP. Os técnicos elaboraram uma lista contendo 10 perguntas abrangendo os sistemas carregados no DW.

Para avaliar a utilidade do componente OLAM foram elaboradas, com o auxílio de técnicos da saúde, algumas questões que pudessem ser respondidas por meio das atividades de mineração de dados, implementadas no MINERSUS. Os resultados produzidos pelas técnicas de mineração foram analisados a partir dos dados de entrada, os quais foram inseridos em planilhas e tabulados adequadamente para comprovação dos resultados. Todas as perguntas foram respondidas mediante a integração dos componentes OLAP e OLAM, ou seja, desenvolvendo inicialmente um relatório OLAP e aplicando as técnicas de mineração sobre os dados resultantes desse relatório. As perguntas desenvolvidas para a avaliação das atividades de mineração estão descritas na Tabela 3.

Tabela 3. Perguntas para a avaliação do componente OLAM. ***Table 3.*** *Questions for OLAM component evaluation.*

	QUESTÕES
	Agrupamento
Q1	Quais são os grupos de capítulos da CID formados em função dos óbitos?
Q2	Quais são os grupos de municípios formados em função da taxa de mortalidade infantil?
Q3	Quais são os grupos de municípios formados em função do gasto na campanha contra a catarata?
	Associação
Q4	Há alguma relação entre óbitos causados por cânceres e as dimensões faixa etária, raça e sexo?
Q5	Há alguma relação entre óbitos causados por doenças do sistema circulatório e as dimensões faixa etária, raça, sexo e grau de instrução?
Q6	Considerando que a gripe é uma patologia que ocorre com maior freqüência no inverno, e conseqüentemente eleva o custo hospitalar, quais são as outras patologias que apresentam uma variação temporal similar à gripe?
Q7	Quais patologias, pertencentes ao capítulo CID I, apresentam similaridade na variação espacial em função das Regionais de Saúde, quando analisadas pela quantidade de internações?
	Classificação
Q8	É possível estabelecer um perfil para mães cujos filhos nasceram com Síndrome de Down, considerando as dimensões faixa etária, raça e escolaridade?
Q9	É possível estabelecer um perfil para os óbitos decorrentes de Diabetes, considerando as dimensões faixa etária, raça, grau de instrução, estado civil e sexo?

A avaliação da usabilidade foi realizada por meio de uma pesquisa de campo que avaliou a interação do usuário com a ferramenta. Foram elaboradas questões sobre saúde pública, com auxílio de técnicos da saúde, para serem respondidas pelo MINERSUS, cujo funcionamento foi demonstrado previamente aos voluntários. Após a demonstração, os voluntários interagiram com o componente analítico do MINERSUS e o pesquisador registrou o desempenho do voluntário.

Foram elaboradas duas perguntas para cada atividade de mineração de dados, exigindo do voluntário o uso de diferentes sistemas do DATASUS e permitindo a avaliação do melhor desempenho. Os voluntários também expressaram a sua opinião sobre a usabilidade do ambiente.

O grupo de voluntários foi constituído por três categorias de usuários, cada uma contendo cinco pessoas. As categorias são: profissionais da saúde pública; técnicos em informática, e administradores de empresa.

Os administradores de empresas representam uma categoria de usuário importante para a avaliação, pois correspondem ao público alvo do ambiente: os tomadores de decisão. Este grupo foi constituído por gerentes e diretores de empresas, com um mínimo de cinco anos na função, pertencendo a diversos segmentos da economia. O grupo dos técnicos em informática foi constituído por profissionais que trabalham no suporte de *software* ao usuário final. Neste grupo foram incluídos profissionais com conhecimentos básicos, porém sem experiência em ferramentas OLAP e técnicas de mineração de dados. O grupo dos profissionais da saúde pública foi constituído por usuários que exercem alguma atividade na saúde pública por mais de três anos e com alguma experiência nos sistemas de informação do SUS.

A demonstração da ferramenta para os voluntários foi realizada em 40 minutos, divididos em duas partes de 20 minutos. A primeira foi dedicada à conceituação das tecnologias, e na segunda foi explicado o processo para elaboração de um relatório OLAP e para a criação e execução dos modelos de mineração de dados. Os voluntários estavam autorizados a esclarecer dúvidas ou solicitar o auxílio do pesquisador durante a interação com o MINERSUS, porém, restritos a perguntas que não contribuíssem diretamente para a criação do relatório ou do modelo de mineração.

O desempenho do voluntário foi registrado de acordo com as alternativas: respondeu sem auxílio; respondeu com auxílio sobre o negócio; respondeu com auxílio sobre o modelo de dados; respondeu com auxílio sobre conceitos de mineração de dados; respondeu com auxílio sobre aspectos operacionais da ferramenta; não conseguiu responder. A opinião dos voluntários sobre a facilidade de uso da ferramenta foi registrada por eles, de acordo com dois itens: a facilidade para elaborar os modelos de mineração e a facilidade para entender os resultados apresentados. A resposta foi restrita a um dos valores: muito fácil, fácil, médio, difícil e muito difícil.

Resultados

O MINERSUS foi implementado integralmente, incluindo a análise e documentação das bases de dados dos sistemas do SUS, a modelagem do DW, a carga dos dados, a implementação dos componentes ETL e dos módulos de processamento analítico OLAP e OLAM.

Foram carregados no DW os dados referentes ao Estado de São Paulo dos sistemas: CNES, SIA, SIH, SIM e SINASC. O período considerado foi de um ano (2005) para os sistemas: SIA, SIH e CNES; e de cinco anos (2000 até 2004) para o SIM e SINASC. A quantidade de registros incluídos foi: CNES = 2.504.411; SAI = 22.211.752; SIH = 21.035.392; SIM = 1.440.908; SINASC = 3.790.279 e Auxiliares = 928.288.

A infra-estrutura para avaliação foi constituída por um servidor com um processador AMD Athlon™ FX Dual Core 2 GHz, 4 GB de RAM, 4 HD de 250 GB instalados em RAID-5, sistema operacional Windows® 2003 Server 64 bits e sistema gerenciador de banco de dados MS-SQL Server 2005.

As ferramentas (ETL, OLAP e OLAM) foram implementadas em MS-Visual Basic 6.0, contemplando todas as funcionalidades propostas para o ambiente. A ferramenta OLAM implementada atende 73% dos itens estabelecidos na metodologia de avaliação de uma ferramenta de mineração de dados proposta por Collier *et al*. (1999).

Para avaliação da utilidade do componente OLAP do MINERSUS, foram desenvolvidos relatórios que atendiam 225 das 242 planilhas disponíveis (93%). Não foi possível atender 100% porque algumas planilhas continham dados provenientes de sistemas do DATASUS que não estão no escopo do DW. Além dos relatórios, o ambiente foi capaz de produzir informação para responder satisfatoriamente a todas as perguntas contidas na lista elaborada pelos usuários experientes da SES-SP.

Por meio do componente OLAM do MINERSUS foi possível extrair informação suficiente para responder a todas as questões definidas na Tabela 3, avalian-

do a utilidade de cada atividade de mineração de dados contida no componente.

Para a avaliação da atividade de agrupamento foram aplicadas as duas técnicas de agrupamento, *K-means* e ARMD. O parâmetro Quantidade de Grupos foi configurado com os valores 3, 5 e 10, resultando em 3 experimentos para cada pergunta. O ARMD, que é apropriado para isolar os valores dissonantes, apresentou a melhor disposição espacial entre os grupos criados, inserindo no mesmo grupo apenas elementos que possuem as menores distâncias em relação aos seus vizinhos (Figura 7). Para ambas as técnicas, o grupo A concentrou o maior número de elementos.

Para avaliação da coerência dos grupos formados, foram realizados cálculos com os dados de entrada. Foram calculadas as distâncias entre os elementos dos grupos formados, para o ARMD, e a distância entre os elementos e a média do grupo, para o *K-Means*. Em ambos os casos o elemento extremo foi movido para o grupo vizinho, e as distâncias e médias foram recalculadas para verificar se os novos grupos apresentavam distâncias menores que aquelas apresentadas anteriormente. Em todos os casos, os elementos extremos dos grupos apresentaram distâncias menores para o seu grupo do que para o grupo vizinho, constatando a coerência dos grupos formados.

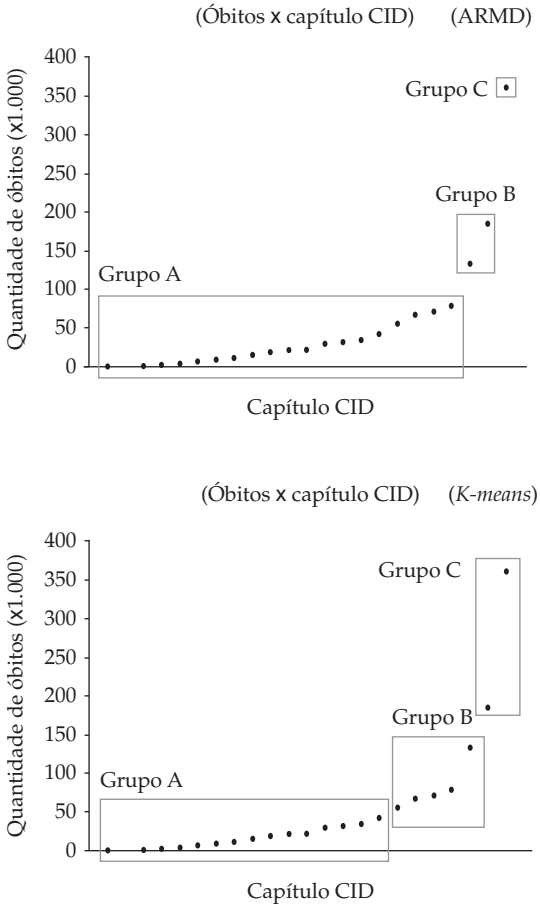

Figura 7. Gráficos agrupamento: usando ARMD; e usando *K-means*. ***Figure 7.*** *Clustering graphics: using CMDR; and using K-means.*

Para a avaliação da atividade de associação, as questões Q4 e Q5 foram respondidas pela técnica tradicional *Apriori*, e as questões Q6 e Q7 pela AVTE.

A avaliação das regras estabelecidas foi realizada por meio do conjunto de dados de entrada. A análise ficou restrita apenas às cinco regras com maior grau de confiança. Os dados referentes aos elementos contidos nas regras analisadas foram inseridos em uma planilha Excel® e foram calculados o suporte e a confiança para cada elemento. Os valores calculados foram comparados com os valores produzidos pela ferramenta e confirmaram a coerência das regras avaliadas.

Para a questão Q4, foram atribuídos os valores 0,1 para o parâmetro Suporte e 10 para a Confiança. Com estes parâmetros foram produzidas 1.269 associações, das quais 116 atingiram os valores definidos para suporte e confiança. Para a questão Q5 foram estabelecidos os mesmos parâmetros e foram produzidas 1.011 associações, das quais 151 atingiram o suporte e a confiança.

As questões Q6 e Q7 foram elaboradas com o objetivo de avaliar o mecanismo AVTE. Os parâmetros utilizados, em ambos os casos, foram: Faixas de Variação = 20, Limite de Variação = 50 e Grau de Confiança = 70. Para Q6, a ferramenta produziu 1.270 associações, das quais 200 atingiram o grau de confiança ≥ 70. Para Q7 a ferramenta produziu 2.071 associações, das quais apenas 23 atingiram o grau de confiança > 70. Reduzindo a confiança para 60, o número de associações que atinge este valor sobe para 67.

A avaliação da atividade de classificação (Q8 e Q9) foi realizada de forma semelhante à avaliação da atividade de associação; as respostas obtidas pela ferramenta foram analisadas a partir dos dados originais para constatação de sua coerência. Foi verificada a coerência para ambas as questões.

Os resultados da pesquisa de campo realizada para avaliação da usabilidade do MINERSUS estão resumidos na Tabela 4. A primeira coluna da tabela mostra os itens da avaliação, a segunda e terceira colunas mostram a quantidade de voluntários que se enquadram no item, em números absolutos e percentuais, respectivamente.

Tabela 4. Resumo dos resultados da avaliação de usabilidade. ***Table 4.*** *Usability evaluation results.*

Estatísticas sobre a interação do usuário com a ferramenta	Voluntários	%
Conseguiram responder a todas as questões	13	87
Não conseguiram responder a todas as questões	2	13
Concluíram as tarefas sem nenhuma ajuda	9	60
Concluíram as tarefas com algum tipo de ajuda	6	40
Tipo de Ajuda solicitada		
Sobre o negócio	1	7
Sobre o modelo de dados	1	7
Sobre o conceito de mineração de dados	2	13
Sobre aspectos operacionais	2	13
Opinião do voluntário sobre a facilidade para criar os modelos		
Muito Fácil	7	47
Fácil	8	53
Opinião do voluntário sobre a facilidade para entender resultados		
Muito Fácil	8	53
Fácil	7	47

Discussão

Os componentes propostos para o MINERSUS são dotados de características adequadas ao contexto dos sistemas de informação do SUS, como por exemplo, os recursos para *downloads* e extração de arquivos que, embora sejam tarefas simples, consumiriam aproximadamente três horas de trabalho em cada carga. Outros recursos, tais como análise de estrutura, análise de conteúdo, versionamento e registro de *log* integrado com manutenção, contribuem para garantir a qualidade dos dados carregados e, conseqüentemente, da informação produzida.

O componente analítico foi implementado privilegiando a usabilidade. O uso de assistentes com textos explicativos para conduzir as ações do usuário, mostrou-se uma estratégia decisiva para a facilidade de uso da ferramenta. A pesquisa de usabilidade mostrou que 13 pessoas, mesmo sem conhecimentos aprofundados em estatística e sem treinamento adequado, conseguiram interagir com o MINERSUS e produzir informação para responder a perguntas específicas sobre a saúde pública. Foi observado que os dois voluntários que não conseguiram responder a todas as questões não apresentaram dificuldades para interagir com a ferramenta, mas sim em compreender o conceito e a aplicação das atividades de mineração. Isto sugere que, para os treinamentos deste tipo de ferramenta, é muito importante abordar detalhadamente os conceitos e a aplicabilidade das técnicas de mineração de dados. Todos os voluntários, inclusive aqueles que não conseguiram realizar todas as tarefas, definiram o processo de criação dos modelos e a visualização dos resultados como muito fácil (53%) ou fácil (47%).

O uso de assistentes reduz a possibilidade do usuário produzir informações inadequadas, pois a seqüência de ações é apresentada passo a passo, exigindo apenas uma escolha do usuário em cada tela. Isto facilitaria a disseminação do MINERSUS para outros usuários, além de gestores da saúde pública, como pesquisadores, estudantes e até mesmo cidadãos comuns.

Uma das grandes contribuições proporcionadas pelo MINERSUS é a agilidade na elaboração de relatórios integrando dados dos diferentes sistemas da saúde pública. No contexto atual, a produção de um relatório contendo dados de diferentes sistemas do SUS demanda um esforço de cinco horas, em média, de um profissional com habilidade num conjunto de ferramentas necessárias à obtenção, extração e integração de dados. Com o MINERSUS, o próprio gestor da saúde consegue obter tais informações em alguns minutos.

A integração das tecnologias OLAP e OLAM no MINERSUS é bem mais abrangente do que a simples existência das duas tecnologias numa mesma plataforma. Deste modo as atividades de mineração de dados complementam o processo analítico, dando ao gestor de saúde mais informação que a produzida pela tecnologia OLAP. O diferencial entre o MINERSUS e algumas ferramentas sofisticadas para mineração de dados, como *Megapunter*, *Angoss*, *Weka* e outras, é a abrangência. Tais ferramentas são destinadas exclusivamente à mineração de dados, enquanto o MINERSUS permite desde a emissão de um simples relatório até a detecção de padrões por meio das técnicas de mineração, e tudo isto num fluxo gradativo e contínuo, sem a necessidade de um profissional de informática para preparar os dados ou configurar os modelos de mineração de dados.

Conclusão

Foi definido, implantado e avaliado um ambiente computacional para extração de informação analítica a partir da mineração das bases de dados do SUS. Neste ambiente foi definido e implementado um DW, um componente para a carga e um componente para produção de informação analítica, através das tecnologias OLAP e OLAM. Foram implementadas três ati-

vidades de mineração, contendo alguns mecanismos adicionais como ARMD e AVTE.

A utilidade do MINERSUS foi avaliada por meio da aplicação de alguns exemplos e a usabilidade por meio de uma pesquisa de campo. Em ambas as avaliações os resultados foram positivos, confirmando a capacidade do ambiente em extrair informações úteis à gestão da saúde pública e a facilidade de uso proporcionada ao gestor.

Alguns assuntos merecem aprofundamento em futuras pesquisas, como a inclusão de outras atividades e técnicas de mineração de dados, um mecanismo mais eficiente para detecção do versionamento de dimensões e um método para detecção automática dos parâmetros ideais para os algoritmos de mineração de dados.

O trabalho deixa as duas contribuições esperadas: a metodologia para a construção deste ambiente computacional e o ambiente implantado e disponível para uso.

Esperamos que o uso do MINERSUS seja difundido e que possa contribuir significativamente para o aumento da qualidade, eficácia e eficiência da gestão da saúde pública.

Referências

BALLARD, C.; HERREMAN, D.; SCHAU, D.; BELL, R.; KIM, E.; VALENCIC, A. **Data Modeling Techniques for Data Warehousing**. San Jose: RedBooks IBM Corporation, 1998.

BELLAZZI, R.; ZUPAN, B. Predictive data mining in clinical medicine: current issues and guidelines. **International Journal of Medical Informatics,** v. 77, n. 2, p. 81-97, 2008.

BERNDT, D. J.; HEVNER, A. R.; STUDNICKI, J. CATCH/IT: A data warehouse to support comprehensive assessment for tracking community health. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION (AMIA) ANNUAL SYMPOSIUM, 7-11 Nov, 1998. **Proceedings…** Orlando, 1998, p. 250-254.

BERNDT, D. J.; HEVNER, A. R., STUDNICKI, J. The Catch data warehouse: support for community health care decision-making. **Decision Support Systems**, v. 35, n. 3, p. 367-384, 2003.

BERSON, A.; SMITH, S. J. **Data Warehousing, Data Mining, & OLAP**. New York: McGraw-Hill, 1997.

BRASIL. **Constituição da República Federativa do Brasil**. Brasília: Senado Federal.

BREEN, C.; RODRIGUES, L. M. Implementing a data warehouse at Inglis Innovative Services. **Journal of Healthcare Information Management**, v. 15, n. 2, p. 87-97, 1988.

BROSSETTE, S. E.; SPRAGUE, A. P.; HARDIN, J. M., WAITES, K. B.; JONES, W. T.; MOSER, S. A. Association rules and data mining in hospital infection control and public health surveillance. **Journal of the American Medical Informatics Association**, v. 5, n. 4, p. 373-38, 1998.

CHAE, Y. M.; HO, S. H.; CHO, K. W.; LEE, D. H.; JI, S. H. Data mining approach to policy analysis in health insurance domain. **International Journal of Medical Informatics**, v. 62, n. 2-3, p. 103-111, 2001.

COLLIER, K.; CAREY, B.; SAUTTER, D.; MARJANIEMI, C. A Methodology for Evaluating and Selecting Data Mining Software. In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 32, 1999, Maui. **Proceedings…** Maui, 1999, v. 6, 11 p.

DATASUS. **Departamento de Informática do SUS.** Disponível em: <www.datasus.gov.br>. Acesso em: 24 abr. 2005.

DEJESUS, E. X. Disease management in a warehouse: data warehouse technology makes a good fit for disease management programs. **Healthcare Informatics**, v. 16, n. 9, p. 33-39, 1999.

EDER, J.; KONCILIA, C. Changes of dimension data in temporal data warehouses. **Lecture Notes in Computer Science**, v. 2114, p. 284-293, 2001.

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. **SIGKDD Explorations**, v. 1, n. 1, p. 20-33, 1999.

HAN, J. Towards on-line analytical mining in large databases. **ACM SIGMOD Record**, v. 27, n. 1, p. 97-107, 1998.

HAN, J.; CHIANG, J. Y.; CHEE, S.; CHEN, J.; CHEN, Q., CHENG, S.; GONG, W.; KAMBER, M.; KOPERSKI, K.; LIU, G.; LU, Y.; STEFANOVIC, N.; WINSTONE, L.; XIA, B. B.; ZAIANE, O. R.; ZHANG, S.; ZHU, H. DBMiner: a system for data mining in relational databases and data warehouses. In: CONFERENCE OF THE CENTRE FOR ADVANCED STUDIES ON COLLABORATIVE RESEARCH, 10-13 nov., 1997, Toronto. **Proceedings…** IBM Press, 1997, p. 8-19.

HAN, J.; KAMBER, M. **Data Mining:** Concepts and Techniques. 2 ed. San Franciso: Morgan Kaufmann, 2006.

INMON, W. H. **Como Construir o Data Warehouse**. 2 ed. Rio de Janeiro: Campus. 1997.

KIMBALL, R. **Data Warehouse Toolkit**. São Paulo: Makron Books, 1998.

MCGARRY, K. A survey of interestingness measures for knowledge discovery. **The Knowledge Engineering Review**, v. 20, n. 1, p. 39-61, 2005.

RAMICK, D. C. Data warehousing in disease management programs. **Journal of Healthcare Information Management**, v. 15, n. 2, p. 99-105, 2001.

SANTOS, R. S.; GUTIERREZ, M. A.; TACHINARDI, U.; FURUIE, S. S. Projeto de data warehouse para a saúde pública. In: CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAUDE, 7-10 nov., 2004, Ribeirão Preto. **Anais...** Ribeirão Preto, 2004, p. 131-136.

SANTOS, R. S.; ALMEIDA, A. L.; TACHINARDI, U.; GUTIERREZ, M. A. Data warehouse para a saúde pública: estudo de caso SES-SP. In: CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAUDE, 10, 14-18 out, Florianópolis, 2006. **Anais...** Florianópolis, 2006, p. 53-58.

SOARES, C.; COSTA, J.; BRAZDIL, P. B. A simple and intuitive measure for multicriteria evaluation of classification algorithms. In: WORKSHOP ON META-LEARNING: BUILDING AUTOMATIC ADVICE STRATEGIES FOR

MODEL SELECTION AND METHOD COMBINATION, ASSOCIATED WITH CONF. ECML, 30 jun-02 jul, 2000. **Proceedings…** Barcelona, 2000, p. 87-96.

TANG, Z.; MACLENNAN, J. **Data Mining with SQL Server 2005**. Indianapolis: Wiley Publisher, 2005.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. San Diego: Academic Press, 1999.

YANG, W.; HWANG, S. A process-mining framework for the detection of healthcare fraud and abuse. **Expert Systems with Applications**, v. 31, n. 1, p. 56-68. 2006.